KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Technology
Department of Electrical Engineering

Spatial audio analysis with constrained microphone setups in adverse acoustic conditions

Elisa Tengan

Supervisor:
Prof. dr. ir. T. van Waterschoot

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Technology (PhD)

February 2024

Spatial audio analysis with constrained microphone setups in adverse acoustic conditions

Elisa TENGAN

Examination committee:
Prof. dr. ir. M. Vanierschot, chair
Prof. dr. ir. T. van Waterschoot, supervisor
Prof. dr. ir. M. Moonen
Prof. dr. ir. P. Vandewalle
Prof. dr. ir. P. Karsmakers
Dr. ir. Ann Spriet
(Goodix Technologies Belgium B.V.)
Prof. dr. Andreas Jakobsson
(Lund University)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Technology (PhD)

February 2024

© 2024 KU Leuven – Faculty of Engineering Technology
Uitgegeven in eigen beheer, Elisa Tengan, Kasteelpark Arenberg 10, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Preface

While writing my thesis, I often wondered what this preface would look like. I initially thought that the words would flow effortlessly after submitting the first version of the manuscript and taking a break. I now realize how mistaken I was.

Although feeling stuck while writing has not been unusual for me at all, the reasons behind it now feel much different than from when I was writing a paper, or a thesis chapter. This is, perhaps, because of the immense amount of support I had during my PhD journey, such that expressing my gratefulness in only a few pages becomes overwhelming on its own. Nevertheless, I will give it a try. Most likely, as it happens to research publications, the ideas shared here will be further extended into more detail later on.

First and foremost, I would like to thank my supervisor, Prof. Toon van Waterschoot. I vividly recall our first meeting in person, on my first day at ESAT. You were getting ready to go for a lecture, but you still made sure that, before you left, I knew where the *good* coffee machine was. As trivial as this gesture may seem, it convinced me that I had made the right decision to come to Belgium for my PhD, and that I was truly welcome in the team. Over the five years of this whole journey, your guidance has greatly contributed to my growth, not only as a researcher, but also as a person. Having a PhD advisor genuinely invested in a student's well-being, as you are, is somewhat of a rare privilege, and I thank you for believing in me, especially when I did not believe in myself.

I would like to extend my gratitude to two other key contributors to the work presented in this thesis: Thomas Dietzen and Filip Elvander. Your indispensable roles pushed this research far beyond what I could have achieved on my own. Now that I reflect on my path, I realize that during the moments when I felt as if I couldn't go any further, when all my energy was drained and my brain made it difficult to see a positive outlook, you, alongside Toon, were the ones helping me get back on my feet and reach the finish line of this marathon. Witnessing

your subsequent success fills me with joy, and I consider myself incredibly lucky to have collaborated with such outstanding researchers.

I thank all my colleagues at ESAT for their support, as well as for all the fun moments we shared together. Thank you for our technical discussions, Zoom coffee breaks during COVID, the summer barbecues, and overall your company during all these years. A special thanks goes to Randy, my office mate, with whom I've had many adventures conducting measurements, and Martin, who started and ended this journey with me. Thank you for all our insightful discussions about research and life!

I also thank the students at Groep T, who not only showed me how much I actually enjoy teaching, but also taught me the values of perseverance, patience, and the acknowledgment of individual differences.

The 'Boulder One' climbing gym in Herent holds a special place in my heart as my second home in Belgium. To the staff and my bouldering crew — thank you for all your moral support, for understanding that 'tall beta' doesn't work for me, and for countless inspiring climbing sessions. Overall, thank you for preserving the welcoming environment that we seek within our climbing community. May we still have many other moments of hard climbing being paid off (most likely followed by injuries)!

A heartfelt thanks to my friends, Daniel, Dilara, and Thierry, for welcoming me into Belgium when I was all alone, and sharing significant moments of our lives together. Your support has been invaluable during this stage of my life, and I am truly grateful for having met you.

I thank the Brazilian friends I have made in Belgium: João, Léo, Mansour, Marcela and Walter. I am so grateful for knowing that I can always complain about the Belgian weather with you! Jokes aside, thank you for your support throughout these years, for reminding me of where we come from and the challenges we all faced as Brazilian expats.

I would like to thank Charlene, for all the knowledge on human psychology shared with me throughout the past few years. Thank you for being a great listener, cheerleader, and teacher. Above all, thank you for helping me see that I am enough, I am not alone, and I deserve the things I have achieved so far.

To Laísa, Liane Sumi, Maitê and Marcela — the best *amiguinhas* I could ever wish for — we have known each other for at least twenty years, and we have had the chance to grow up together and see each other following very different paths in life. We are nothing like the kids we were before, yet our love for each other only grows. Thank you for the profoundly positive impact you had in my life through all these years, and thank you for being amazing, inspiring women,

whom I am very proud to call my childhood friends!

To Amanda, another amazing and inspiring woman — our friendship, cultivated since high school, has evolved into a deep, special bond. Thank you for your unwavering support, our gastronomic adventures, and your wisdom!

I would like to thank my 'second mom', Paoline, and my 'brother', Ivan, for their support and for letting me be part of their family. To my parents — Hiromi and Flaviano — thank you for always believing in me and for supporting me, even when you may not understand some of my choices and decisions (to be honest, I don't always understand them either). I feel very grateful to know that no matter the distance, you are always there for me.

Last but not least, I would like to thank my family in Leuven, an unexpected and invaluable addition to my life during this PhD journey: Akash, Miriam, Mohit and Santiago. It is incredibly difficult for me to put into words how important you are in my life. We come from very different backgrounds, such that five years ago, I would have never expected us to become so close. With time, however, it became clear to me that your kindness, truthfulness, trust, and empathy, are the most cherished treasures I could ever ask for. Thank you for always standing by my side, understanding me, being my official COVID bubble, and, above all, for the gift of your enduring friendship.

Abstract

The spatial information contained in recorded microphone signals has a fundamental role in the characterization of acoustic environments. Furthermore, it can be exploited in digital audio processing frameworks for enhancing a desired source signal, benefiting applications such as hands-free telephony, hearing devices, human-machine interfaces and acoustic monitoring systems. Despite its acknowledged relevance, accurately estimating and effectively employing spatial information can present significant challenges, due to the diverse acoustic conditions of practical scenarios or limitations imposed by microphone setups.

This thesis focuses on the development and evaluation of audio processing methods for estimating and applying spatial information, while addressing a selection of challenges encountered in different applications. The estimation of spatial information is, in this thesis, limited to the problem of source localization. Firstly, two approaches for single-channel source localization are presented, motivated by the limitations in spatial audio analysis capabilities imposed by single-microphone setups, or devices in which simultaneous access to multiple microphone signals is unreliable. Both approaches estimate the direction of arrival (DOA) and power spectral density (PSD) of stationary point sources by using a single, rotating, directional microphone. The solution to a group-sparse regularized optimization problem is used for estimating direction-dependent PSD values relative to a given angular dictionary, and locating peaks in the estimated PSD vector. The methods' performance is evaluated through a series of simulations in which different setup conditions are considered, ranging from different types of model mismatch to variations in the acoustic scene and microphone directivity pattern.

Building on the insights gained from evaluating the single-channel localization methods, a multi-channel source localization method is also presented in this thesis. This method estimates the DOA of multiple broadband sound sources through a similar group-sparse optimization problem, modeling an observed broadband steered response power (SRP) map as a linear combination of

PSDs instead. Simulation results demonstrate that the proposed method outperforms conventional methods in certain multi-source scenarios and maintains comparable performance to others while being more computationally efficient. Moreover, it shows superior performance in cases with closely located sources.

After addressing the problem of source localization, this thesis shifts its focus to exploiting spatial information for signal enhancement. The particular task of of speech enhancement with a microphone array embedded in an unmanned aerial vehicle (UAV) is considered, which entails highly adverse acoustic conditions, involving high levels of ego-noise. Firstly, an experimental methodology is presented for measuring ego-noise emissions. Then, the feasibility of employing spatial filtering-based methods for speech enhancement in such a challenging scenario is investigated. A speech enhancement method using ego-noise references is presented, based on the prior-knowledge multi-channel Wiener filter (PK-MWF) and speech presence probability (SPP) estimates. Additionally, the method's development is extended to explicitely incorporate estimated DOA information on the target speaker, yielding an alternative implementation. The proposed approaches are evaluated over experimental recordings obtained with a drone operating at constant thrust, in a simulated hovering condition, and results indicate the importance of reliable speech activity estimates in the performance of spatial filtering-based methods such as the PK-MWF. It is shown that employing the DOA-based implementation of the PK-MWF filter significantly improves the signal-to-noise ratio (SNR) of the filtered signal when compared to the result obtained when employing the PK-MWF with SPP estimated from one of the embedded microphones, whereas the perceptual improvement is compromised. Furthermore, it is shown that combining both DOA and SPP-based implementations of PK-MWF can improve performance in terms of perceptual speech quality.

This thesis also presents research valorisation plan, indicating how the research outcomes can be transferred into varying application contexts, while also outlining the potential socio-economic value they could generate in an industrial setting.

Finally, the thesis is concluded with a summary of the contributions presented and suggestions for future research, while considering the essential validation aspects required.

Beknopte samenvatting

De ruimtelijke informatie in opgenomen microfoonsignalen heeft een fundamentele rol bij de karakterisering van akoestische omgevingen. Bovendien kan het worden benut bij digitale audioverwerking raamwerken voor het verbeteren van een gewenst bronsignaal, wat toepassingen als deze ten goede komt als handsfree telefonie, hoortoestellen, mens-machine-interfaces en akoestische monitoring systemen. Ondanks de erkende relevantie, accuraat inschattend en effectief Het gebruik van ruimtelijke informatie kan aanzienlijke uitdagingen met zich meebrengen, vanwege de divers akoestisch voorwaarden van praktische scenario's of beperkingen opgelegd door microfoonopstellingen.

Dit proefschrift richt zich op de ontwikkeling en evaluatie van audioverwerking methoden voor het schatten en toepassen van ruimtelijke informatie, terwijl een selectie van uitdagingen wordt aangepakt die men in verschillende landen tegenkomt toepassingen. De schatting van ruimtelijke informatie is, in dit proefschrift, beperkt tot het probleem van bronlokalisatie. Ten eerste twee benaderingen voor single-channel bronlokalisatie worden gepresenteerd, gemotiveerd door de beperkingen in ruimtelijke audio analysemogelijkheden opgelegd door opstellingen met één microfoon, of apparaten waarin gelijktijdige toegang tot meerdere microfoonsignalen is onbetrouwbaar. Beide benaderingen schatten de richting van aankomst (DOA) en spectrale vermogensdichtheid (PSD) van stationair puntbronnen met behulp van een enkele, roterende, richtmicrofoon. De oplossing voor een groep-sparse geregulariseerd optimalisatieprobleem wordt gebruikt voor het schatten richtingsafhankelijke PSD-waarden ten opzichte van a gegeven hoekwoordenboek, en het lokaliseren van pieken in de geschatte PSD-vector. De De prestaties van methoden worden geëvalueerd door middel van een reeks simulaties met verschillende instellingen Er wordt rekening gehouden met omstandigheden, variërend van verschillende soorten modelmismatches variaties in de akoestische scène en het richtingspatroon van de microfoon.

Voortbouwend op de inzichten die zijn verkregen door de evaluatie van lokalisatie via één kanaal methoden wordt in dit proefschrift ook een multi-channel

bronlokalisatiemethode gepresenteerd. Dit methode schat de DOA van meerdere breedbandgeluidsbronnen via a soortgelijk groep-sparse optimalisatieprobleem, waarbij een waargenomen breedbandgestuurde modellering wordt gegeven responsvermogen (SRP) in kaart gebracht als een lineaire combinatie van PSD's. Simulatie resultaten aantonen dat de voorgestelde methode op bepaalde punten beter presteert dan conventionele methoden scenario's met meerdere bronnen en handhaaft vergelijkbare prestaties als andere rekentechnisch efficiënter zijn. Bovendien toont het superieure prestaties in gevallen met dicht bij elkaar gelegen bronnen.

Nadat het probleem van bronlokalisatie is behandeld, verlegt dit proefschrift de focus het benutten van ruimtelijke informatie voor signaalverbetering. De bijzondere taak van van spraakverbetering met een microfoonarray ingebed in een onbemande luchtvoertuig (UAV) wordt overwogen, wat brengt zeer ongunstige akoestische omstandigheden met zich mee, met hoge niveaus van ego-ruis. Ten eerste wordt een experimentele methodologie voor het meten gepresenteerd ego-geluidsemissies. Vervolgens wordt de haalbaarheid van het gebruik van ruimtelijke filtering onderzocht methoden voor spraakverbetering in een dergelijk uitdagend scenario worden onderzocht. Een spraakverbeteringsmethode met behulp van ego-ruisreferenties wordt gepresenteerd, gebaseerd op de voorkennis meerkanaals Wiener-filter (PK-MWF) en spraakaanwezigheid waarschijnlijkheidsschattingen (SPP). Bovendien is de methode De ontwikkeling wordt uitgebreid om expliciet geschatte DOA-informatie op te nemen doelspreker, wat een alternatieve implementatie oplevert. De voorgestelde benaderingen worden experimenteel geëvalueerd opnames verkregen met een drone die met constante stuwkracht werkt, in een gesimuleerde vorm zwevende toestand, en de resultaten wijzen op het belang van betrouwbare spraak activiteitsschattingen bij de uitvoering van op ruimtelijke filtering gebaseerde methoden zoals de PK-MWF. Er wordt aangetoond dat gebruik wordt gemaakt van de DOA-gebaseerde implementatie van het PK-MWF-filter verbetert de signaal-ruisverhouding (SNR) van het gefilterde signaal aanzienlijk in vergelijking met de resultaat verkregen bij gebruik van de PK-MWF met SPP geschat op basis van een van de ingebouwde microfoons, terwijl de perceptuele verbetering in gevaar komt. Verder wordt aangetoond dat combineren zowel DOA- als SPP-gebaseerde implementaties van PK-MWF kunnen hiervoor worden gebruikt de prestaties op het gebied van perceptuele spraakkwaliteit verbeteren.

Dit proefschrift presenteert ook een onderzoeksvalorisatieplan, waarin wordt aangegeven hoe het onderzoek zal resulteren kan worden overgedragen naar verschillende toepassingscontexten, terwijl ook waarin de potentiële sociaal-economische waarde wordt geschetst die ze in een industriële omgeving kunnen genereren.

Glossary

Abbreviations and acronyms

ADMM	alternating direction method of multipliers
ASR	automatic speech recognition
CAGR	compound annual growth rate
Ch.	Chapter
cf.	*confer*, compare with, see also
DFT	discrete Fourier transform
DOA	direction of arrival
e.g.	*exempli gracia*, for example
FFT	fast Fourier transform
Fig.	Figure
GCC	generalized cross correlation
GPS	Global Positioning System
GSC	generalized sidelobe canceler
GSPICE	group-sparse iterative covariance-based estimation
HVAC	heating, ventilation and air conditioning

i.e.	*id est*, that is
iSPP	speech presence probability estimated from embedded microphone
MAE	mean absolute error
MUSIC	multiple signal classification
MVDR	minimum variance distortionless response
NMSE	normalized mean squared error
PHAT	phase transform
PK-MWF	prior knowledge multi-channel Wiener filter
PESQ	perceptual evaluation of speech quality
PR	power ratio
PSD	power spectral density
RTF	relative transfer function room transfer function (Ch. 2)
RTLS	real-time location system
SDW-MWF	speech distortion weighted multi-channel Wiener filter
Sec.	Section
SNR	signal-to-noise ratio
SPP	speech presence probability
SRP	steered response power
SRP-PHAT	steered response power with phase transform
STFT	short-time Fourier transform
STOI	short-time objective intelligibility
TDOA	time difference of arrival

UAV	unmanned aerial vehicle
VAD	voice activity detector
VTOL	vertical takeoff and landing
xSPP	speech presence probability estimated from external microphone

Mathematical notation

a (lower case)	a scalar
$\mathbf{a}$ (bold lower case)	a vector
$\mathbf{A}$ (bold upper case)	a matrix
$\mathbf{A}^\top$	the transpose of $\mathbf{A}$
$\mathbf{A}^\mathrm{H}$	the Hermitian transpose of $\mathbf{A}$
$\Re\{\mathbf{A}\}$	the real part of $\mathbf{A}$
$\mathbf{A}^{-1}$	the inverse of $\mathbf{A}$
$[\mathbf{A}]_{m,n}$	the element of $\mathbf{A}$ at index (m, n)
$\mathbb{R}$	the set of real numbers
$\mathbb{C}$	the set of complex numbers
$\mathbf{I}_N$	an $N \times N$ identity matrix
$\mathbf{1}_K$	a K-element column vector of ones
$\|\mathbf{A}\|_\mathrm{F}$	the Frobenius norm of $\mathbf{A}$
$\|\mathbf{a}\|_p$	the l_p-norm of $\mathbf{a}$
$\mathbb{E}\{a\}$	the expectation of a
$\hat{\mathbf{a}}$	an estimate of $\mathbf{a}$
$\mathrm{diag}(\mathbf{a})$	a diagonal matrix with the elements of $\mathbf{a}$ on its diagonal

a^*	the complex conjugate of a
$\mathrm{mod}(a, b)$	a modulo b
$\lceil a \rceil$	the ceiling of a
e	Euler's number
j	the imaginary unit ADMM iteration index (Ch. 2)

List of symbols

d	propeller diameter (Ch. 4)
d_p	Euclidean distance of microphone pair p (Ch. 3)
i	reweighting iteration index (Ch. 2) source index (Ch. 3)
J	number of sources (Ch. 3)
k	discrete frequency index
λ	regularization parameter(Ch. 2) drone pitch (Ch. 4)
λ_l	regularization parameter for l-th group (Ch. 3)
L	number of candidate DOAs number of observation frames (Ch. 5)
m	microphone index (Ch. 3)
n	observation frame index (Ch. 2)
N	number of microphone orientations (Ch. 2) number of SRP samples (Ch. 3)
P	number of sources (Ch. 2) number of microphone pairs (Ch. 3)
p	source index (Ch. 2) microphone pair index (Ch. 3)

r	realization index radius (Ch. 4)
$\alpha(\omega_k)$	weighting factor for frequency ω_k (Ch. 7)
α_p	rotation of microphone pair p (Ch. 3)
$\Gamma(\theta)$	directivity pattern (Ch. 2)
Γ_N	N-element SRP grid (Ch. 2)
γ_n	microphone orientation at frame n (Ch. 2) n-th element of SRP grid Γ_N (Ch. 3)
Θ_L	L-element candidate DOA grid
θ_l	l-th element of candidate DOA grid Θ_L
ϑ_p	DOA of p-th source
ρ	dual update step size (Ch. 2) elevation (Ch. 4, Ch. 6)
$\rho_i^{(r)}$	accuracy score for i-th source and r-th realization (Ch. 3)
$\bar{\rho}$	average accuracy score (Ch. 3)
ω_k	k-th discrete angular frequency

Contents

Abstract v

Beknopte samenvatting vii

Glossary ix

Contents xv

List of Figures xxi

List of Tables xxv

1 Introduction 1

1.1 Source localization 2

1.2 Speech enhancement 6

1.3 Applications and challenges 7

1.4 Contributions 10

1.5 Thesis outline 15

I Source localization 17

2 Single-channel source localization 19

2.1 Introduction 21
2.2 Signal model 23
2.2.1 Rotational microphone signal model 23
2.2.2 Grid-based system of equations 24
2.3 Proposed approaches 26
2.3.1 Group Lasso followed by least squares (GL-LS) 27
2.3.2 Group Lasso with l_1-reweighting (GL-L1) 29
2.4 Implementation 30
2.5 Simulations 35
2.5.1 Performance measures 37
2.5.2 Different grid resolutions 37
2.5.3 Different levels of microphone response mismatch 40
2.5.4 Different levels of reverberation 42
2.5.5 Influence of angular separation and power ratio between two sources 44
2.5.6 Influence of the microphone directivity pattern in the case of three sources 46
2.5.7 Discussion on the performance limitations of GL-LS and GL-L1 51
2.6 Conclusion 52

3 Multi-channel source localization **53**

3.1 Introduction 54
3.2 Signal model 57
3.3 Broadband steered response power (SRP) mapping 58
3.3.1 Broadband SRP formulation 58
3.3.2 Uniform spatial sampling requirements 59
3.4 Proposed method 62

3.4.1 Group-sparse fitting of a broadband SRP map 63
3.4.2 Group square-root Lasso implementation 64
3.5 Simulations 66
3.5.1 Simulation setup 67
3.5.2 Search grids of candidate DOAs 68
3.5.3 Performance evaluation 69
3.6 Results 71
3.6.1 Number of SRP samples and computational complexity 71
3.6.2 Number of averaging STFT cross-correlation frames . . 73
3.6.3 Angular separation between sources 74
3.6.4 Noise level 75
3.6.5 Reverberation 76
3.6.6 Broadband power ratio between two sources 77
3.6.7 Number of sources 79
3.7 Conclusion 80

II Speech enhancement in adverse acoustic conditions 81

4 Experimental setup for measuring UAV noise 83
4.1 Introduction 85
4.2 Measurement setup 86
4.3 Results and discussion 89
4.4 Conclusion 92

5 Speech enhancement with a UAV 95
5.1 Introduction 97
5.2 Signal model 98

5.3 Proposed method 99
5.3.1 Formulation of the multichannel Wiener filter (MWF) . 99
5.3.2 Standard multichannel Wiener filter (MWF) realization 99
5.3.3 Prior knowledge multichannel Wiener filter (PK-MWF) realization 100
5.4 Evaluation 102
5.4.1 Measurement setup 102
5.4.2 Processing 102
5.5 Results 105
5.6 Conclusion 107

6 DOA-informed speech enhancement with a UAV 111

6.1 Introduction 112
6.2 Direction-of-arrival estimation using a microphone array embedded in a UAV 113
6.2.1 Broadband SRP in terms of azimuth and elevation . . . 113
6.2.2 Weighted broadband SRP based on the spectral characteristics of UAV ego-noise 114
6.3 Speech enhancement using ego-noise references and estimated source DOA 116
6.4 Evaluation 117
6.4.1 Processing 118
6.4.2 Results 119
6.5 Conclusion 122

7 Research valorisation plan 125

7.1 Research outcomes 126
7.1.1 Methods 126
7.1.2 Dataset 126

7.2 Research collaborations with industry partners 127
7.2.1 VLAIO R&D Project no. HBC.2017.0358 "SPOTT - Tomorrow's Scalable and PersOnalised advertising Technology, Today". 127
7.2.2 Experimental measurements with an unmanned aerial vehicle (UAV) 129
7.3 Strategic analysis of relevant markets 130
7.4 Validation 132

8 Conclusion **135**
8.1 Summary 135
8.2 Future research 138

A Appendix to Chapter 2 **141**
A.1 DOA estimation comparison with MUSIC 141

B Appendix to Chapter 3 **145**
B.1 GSPICE applied to microphone cross-correlations 145

Bibliography **147**

Curriculum Vitae **173**

List of publications **175**

Acknowledgements **177**

List of Figures

1.1 Illustration of a listener in a diverse acoustic environment, along with the spatial perception of multiple sound sources. 3

1.2 Example of a setup composed of two microphones, capturing the signal emitted by a source in the far field. The inter-microphone distance is denoted by d, and the source's DOA by θ. 4

1.3 Examples of applications of source localization and speech enhancement methods. 8

1.4 Examples illustrating the different aspects composing adverse acoustic conditions which pose challenges in source localization and speech enhancement. 10

1.5 Illustration depicting a directional microphone with a cardioid directivity pattern, rotated across various known directions. . . . 11

2.1 Estimation errors obtained with GL-LS for different grid resolutions and SNRs 38

2.2 Estimation errors obtained with GL-L1 for different grid resolutions and SNRs 39

2.3 Estimation errors obtained with GL-LS and GL-L1, as well as with their modified versions which exclude the nonnegativity constraint over the error term (GL-LS$_0$ and GL-L1$_0$, respectively), while using different microphone directivity patterns to generate the observed data 41

2.4 Illustration of simulated room where the source position is randomly generated within the area represented in gray (at least 10 cm from the wall and farther than the room's critical distance r_c) 43

2.5 Estimation errors obtained with GL-LS and GL-L1 for different reverberation times 44

2.6 Estimation errors obtained with GL-LS for different values of angular separation and power ratio between two sources 46

2.7 Estimation errors obtained with GL-L1 for different values of angular separation and power ratio between two sources 47

2.8 Microphone directivity patterns considered in this study represented in absolute values 48

2.9 Estimation errors averaged over three sources and obtained with GL-LS for different values of angular separation, while using different microphone directivity patterns 49

2.10 Estimation errors averaged over three sources and obtained with GL-L1 for different values of angular separation, while using different microphone directivity patterns 50

3.1 Example of a microphone pair on the horizontal plane with its corresponding inter-microphone Euclidean distance d_p and rotation α_p. 61

3.2 Illustration of a simulated scenario with two sources at a radial distance of 1.5 m from the center of a 6-element uniform circular array. 69

3.3 Average accuracy score $\bar{\rho}$ and number of per-iteration multiplications N_{mult} as a function of N, for $K = 256$ and $L = 72$, with SNR $= 0$ dB and $\text{T}_{60} = 0$ s. 72

3.4 Average accuracy score obtained with GSPICE, SRP-PHAT and the proposed method for different numbers of averaging STFT cross-correlation frames N_f, with $N = 24$ and $N = 72$. 74

3.5 Average accuracy score obtained with GSPICE, SRP-PHAT and the proposed method for different angular separations between sources, with $N = 24$ and $N = 72$. 74

3.6 Average accuracy score obtained with GSPICE, SRP-PHAT and the proposed method for different SNRs, with $N = 24$ and $N = 72$. 75

3.7 Average accuracy score obtained with GSPICE, SRP-PHAT and the proposed method for different reverberation times (T_{60}), with $N = 24$ and $N = 72$. 76

3.8 Average accuracy score per source obtained with GSPICE, SRP-PHAT and the proposed method for different PRs, with $N = 24$ and $N = 72$. 78

3.9 Average accuracy score obtained with GSPICE, SRP-PHAT and the proposed method for different numbers of sources, with $N = 24$ and $N = 72$. 79

4.1 Experimental setup consisting of a hemispherical microphone array and a UAV mounted on a stand. The binaural head depicted in the background has not been used in the processing of signals in the results presented in Sec. 4.3. 87

4.2 Spherical coordinate system, where γ denotes the azimuth, ρ the elevation, and r the radius. 88

4.3 Schematic of the microphone array geometry employed in the experimental setup, with the positive x-axis direction corresponding to 0° azimuth and 0° elevation. 89

4.4 SPL for 10°, −30°, −50° and −70° emission directivity across various rotor speeds. 90

4.5 SPL in one-third octave bands for 0° azimuth, and 10°, −30°, −50° and −70° elevation angles. 91

5.1 Sketches of experimental setup 103

5.2 Experimental setup in semi-anechoic chamber 104

5.3 Illustration of different microphone array configurations considered in the formulation of MWF and PK-MWF. 106

5.4 SNR improvement in function of original SNRs associated to different average rotor speeds. 108

5.5 STOI improvement in function of original STOI values associated to different average rotor speeds. 109

5.6 PESQ improvement in function of original PESQ scores associated with different average rotor speeds. 110

6.1 Estimated ego-noise PSD from recordings made with one of the propeller microphones, for different average rotor speeds. 115

6.2 Weight values for average rotor speed of 5000 rpm. 116

6.3 Spectrograms of signals recorded with the reference channel of the UAV's microphone array, as well as the corresponding outputs of the compared methods, for an average rotor speed of 5000 rpm. 120

6.4 Spectrograms of output signals when employing the different implementations of PK-MWF being compared, for an average rotor speed of 5000 rpm. 121

6.5 Performance improvement in terms of different objective measures associated with different average rotor speeds ($M_{\mathrm{S+N}} = 12$). . . 123

A.1 RMSE of DOA estimation of two stationary point source signals for varying angular separations and different SNRs, obtained with wideband MUSIC and the proposed method. 143

List of Tables

2.1 Main simulation parameters for each subsection 36
2.2 Coefficient values for microphone directivity patterns used . . . 48

3.1 Source signal types. 67
3.2 Main parameters for each set of simulations, divided by subsections 70

6.1 Absolute DOA estimation errors for different average rotor speeds. 119

Chapter 1

Introduction

Sound has a fundamental role in human life, not only by allowing interpersonal communication through speech, but also by serving as a medium in which environmental cues of potential threats, emotions, or even solitude can be conveyed [1, 2].

An important aspect of sound is the spatial information it carries. For instance, identifying the location of auditory events contributes to a listener's understanding of the overall environmental context, as illustrated in Fig. 1.1. In binaural hearing, the spatial perception of sound depends on many factors, such as the person's position, their head orientation, and the overall shape of head and ears [3]. These factors, combined with the acoustic conditions of the environment, result in a specific combination of propagating sound waves exciting the auditory system [3, 4].

The ways in which the spatial characteristics of sound are perceived by the human auditory system have been extensively investigated [3–5]. Furthermore, they have also motivated developments in audio signal processing [6]. There has been extensive research investigating the possibilities of estimating and exploiting spatial information through different audio analysis methods, while considering the tasks of, e.g., sound source localization, source separation and spatial signal enhancement. Advancements in spatial audio analysis techniques have contributed to improving hearing experience [7, 8], as well as to the further development of voice-based human-machine interfaces [9]. Additionally, the opportunities arising from the detection and classification of acoustic events within an environment have also motivated the development of acoustic monitoring frameworks [10, 11].

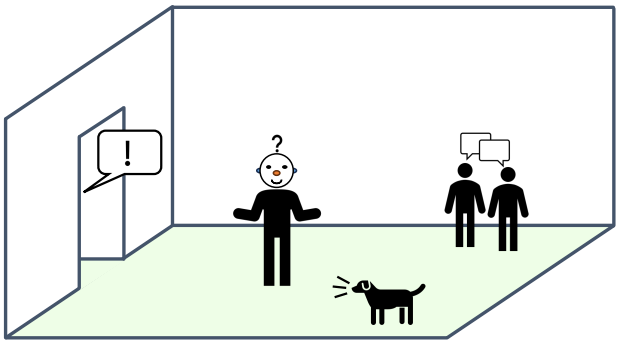
Despite the evident motivations for the use of spatial audio analysis techniques in different applications, accurately estimating and effectively employing spatial information within an audio processing framework can present significant challenges, depending on the acoustic conditions or setup constraints. As part of the continuous effort for addressing some of these challenges, this thesis presents and discusses contributions for a selection of research problems within the subject of spatial audio analysis and enhancement. Specifically, the tasks of source localization and spatial filtering-based speech enhancement are considered.

In this chapter, the thesis is introduced as follows. In Sec. 1.1 and Sec. 1.2, we introduce the research problems of source localization and speech enhancement, respectively, while providing a literature overview. In Sec. 1.3, we briefly discuss some prevalent applications, along with the challenges usually encountered. In Sec. 1.4 we outline the contributions presented this thesis, and in Sec. 1.5, we detail the structure of the remainder of the thesis.

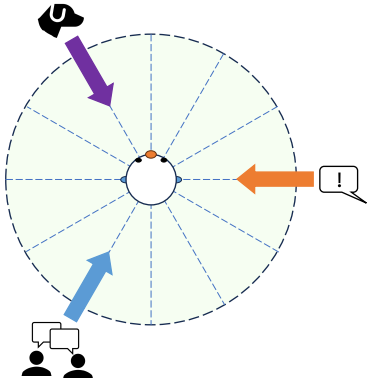
1.1 Source localization

In the field of audio signal processing, sound source localization refers to the task of estimating the location of one or more sound sources, given an acoustic environment and relative to a reference position. Such reference is often defined with respect to the microphone setup being employed [6]. The use of multiple microphones is commonly considered crucial for the task of source localization, since their spatial diversity allows for differences in the microphone signals to be observed and effectively exploited when estimating a source's location [6, 12]. In scenarios where sound sources are in the far field, i.e., at a large enough distance from the microphones, plane-wave propagation is usually assumed for the captured source signals, and the task of localization is reduced to direction-of-arrival (DOA) estimation [6, 7]. An example of a DOA estimation problem for a scenario with a far-field sound source emitting a signal that is captured by a pair of microphones is presented in Fig. 1.2. In the near-field case, i.e, when the sound sources are closer to the microphones, a source's position can also be estimated in terms of its distance from the microphones [6, 13]. Nevertheless, in this thesis, we focus on the former case, and will employ the terms *DOA estimation* and *source localization* interchangeably.

A wide range of source localization methods have been proposed in the literature. Here, we broadly categorize these methods according to their fundamental concepts as follows.



(a) Overview of acoustic environment.



(b) Spatial perception of multiple sound sources and their direction in space.

Figure 1.1: Illustration of a listener in a diverse acoustic environment, along with the spatial perception of multiple sound sources.

- **Beamforming methods** - Many DOA estimation methods rely on the steering of a beamformer over a chosen search grid in order to estimate DOAs by identifying the direction in which the beamformer output is maximized [14]. In the case of the conventional delay-and-sum (DAS) beamformer, this is achieved by applying time shifts to the time-domain array signals to compensate for the propagation delays in the arrival of the source signal at each microphone, for a given set of candidate DOAs [6,14]. Another example of beamforming technique consists in using a minimum-variance distortionless response (MVDR) filter [15, 16]. In the steered response power (SRP) [17, 18] method, an acoustic map is computed as a function of the generalized cross-correlations (GCCs) [19] between different microphone pairs. With formulations in both time and frequency domains,

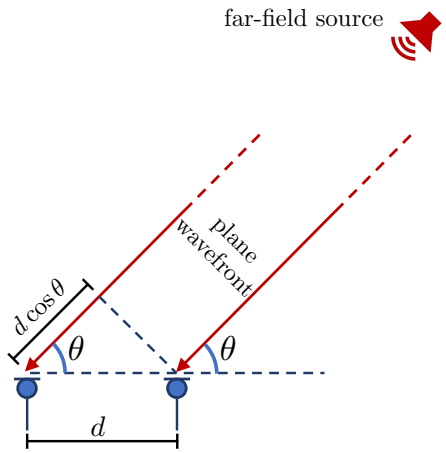


Figure 1.2: Example of a setup composed of two microphones, capturing the signal emitted by a source in the far field. The inter-microphone distance is denoted by d, and the source's DOA by θ.

the conventional SRP method has been extended to improve its robustness in the specific task of speech source localization through the use of the phase transform (PHAT) weighting function in the computations of the GCCs, yielding the well-known SRP-PHAT framework [13, 18].

- **Subspace methods:** Another class of DOA estimation methods corresponds to those relying on a subspace decomposition of a microphone array's signal correlation matrix [14]. In the MUSIC method [20], an eigenvalue decomposition (EVD) is employed to obtain the target signals and noise subspaces, which are in turn used to compute a spatial pseudospectrum for a chosen search grid of candidate DOAs. Hence, the DOA estimation is performed by identifying peaks in the pseudospectrum. The required grid search procedure of MUSIC is considered computationally expensive, such that alternative subspace-based methods have been proposed [21]. For instance, the ESPRIT [22] and root-MUSIC [23] methods allow for search-free DOA estimation, at the expense of having their use constrained to fairly specific microphone array configurations [21]. Both classical ESPRIT and root-MUSIC methods have been later extended to suit more array geometries [24–27].

  Since classical subspace methods such as MUSIC and ESPRIT were initially formulated for the DOA estimation of narrowband signals, it naturally followed that their corresponding broadband versions emerged [28–30]. Additionaly, although classical subspace methods are known to

allow for high resolution localization, their sensitivity to coherence between signals also motivated the development of alternatives when considering reverberant scenarios [31, 32].

- **Maximum likelihood (ML) methods** - This class corresponds to methods in which sources are located based on the maximization of a likelihood function, which is parametrized by the desired target source DOAs, given an assumed signal model and statistical properties of the noise component [33, 34]. Some classical examples of ML-based DOA estimation methods include the alternating projection method (APM) [35], which allows for the localization of multiple narrowband sources through an iterative process of successively estimating each source's DOA. An alternative approach based on APM, focused on broadband signals instead, has been proposed for multi-source DOA estimation in [36], as well as in [37], with the latter being in the spherical harmonics domain. Although ML-based methods for DOA estimation are considered robust in cases of correlated sources or noisy conditions, the inherent high computational complexity (relative to beamforming and subspace methods) has motivated the search for efficient implementations [21, 34].

- **Sparse reconstruction methods** - With the growth of compressed sensing [38, 39], significant efforts have aimed at developing DOA estimation methods in which a sparse representation of the signals observed by the available sensors is exploited [40]. Specifically, the concept of sparse reconstruction relies on the assumption that the observed data, lying on a high-dimensional space, can be well approximated by a reduced-dimensional subspace [40]. Regarding the task of DOA estimation, diverse methods have been proposed. Firstly, sparse DOA estimation methods can differ based on the chosen representation of the observed data [34], which can be based on single or multiple snapshots [41, 42], as well as a beamforming output [43, 44] or a signal covariance matrix [45–47]. Another characteristic differentiating sparse DOA estimation methods is how, or if, a grid of candidate DOAs is employed. While there are many methods that estimate the target source DOAs by identifying the nonzero components of an estimated parameter vector, constructed based on a chosen search grid of DOAs [34, 40, 48–50], other approaches propose to jointly estimate the DOA's off-set with respect to the search grid being used, such that off-grid DOA estimation could be achieved [51–54]. To address grid mismatches, dynamic grid refinement has been suggested as a potential solution [40, 55, 56]. Additionally, in order to circumvent the limitations associated to grid-based approaches, alternative gridless sparse DOA estimation methods have also been proposed [40, 57–59].

1.2 Speech enhancement

In this thesis, speech enhancement refers to the objective of noise reduction over mixed-signal recordings affected by undesired components, in order to recover a clean target speech signal. Although the perceived quality of a restored speech signal is inherently subjective [7, 60], the effectiveness of speech enhancement techniques is commonly evaluated through objective quality measures, due to their practicality [7, 61].

Numerous speech enhancement methods have been proposed in the literature for both single and multi-channel setups [6, 7, 14, 62]. Considering the scope of this thesis, we limit the following overview of the state of the art to spatial filtering-based methods for speech enhancement.

- **Spatial filtering methods:** This class of speech enhancement methods corresponds to *spatially selective* multi-channel filtering techniques. A set of linear filters is applied to multiple microphone signals, such that their resulting combination into a single output aims at enhancing the target signal, while reducing the undesired components. Spatial filters can be differentiated by the optimality criterion used in their design, which can be data-independent or data-dependent. Data-independent spatial filters, such as the delay-and-sum and superdirective beamformers [14], can be implemented while relying on the knowledge of the DOA of the target source, or on the relative transfer function (RTF) from source to microphone array [7, 63]. Alternatively, data-dependent spatial filters commonly rely on the second-order statistics of the observed microphone signals [63, 64]. For instance, the MVDR beamformer [15] minimizes its output power under the constraint that the target signal is not distorted [14]. Another relevant data-dependent method corresponds to the multi-channel Wiener filter (MWF), which considers the minimum mean squared error (MMSE) criterion to estimate the target signal at a reference channel [65]. The MWF can be formulated as an MVDR beamformer followed by a single-channel Wiener post-filtering stage [63], and its use is known to promote greater noise reduction when compared to the MVDR beamformer, at the expense of distorting the estimated output signal [14, 65]. By generalizing the MMSE criterion of the MWF, an explicit trade-off between noise reduction and speech distortion can be achieved with the alternative speech distortion weighted MWF (SDW-MWF) [63, 65–68].

  In practice, the differences between spatial filtering methods proposed for speech enhancement in the literature are not only attributed to the optimality criteria considered, but also to how the required spatial

or statistical information is estimated, when presumed unavailable a priori. Many approaches have been proposed to estimate it based on the recorded microphone signals in a dedicated processing stage of the speech enhancement framework [63]. For instance, voice activity detection (VAD) [69, 70] and speech presence probability (SPP) [71–73] algorithms have been used to classify multiple signal frames (and frequency bins, for SPP) into containing noise only or speech and noise, allowing for the estimation of the signal correlation matrices employed in MVDR and MWF-based approaches [60, 63, 65, 74, 75]. The RTF of the desired target speech signal can be estimated through a generalized eigenvalue decomposition (GEVD) procedure [6, 65, 76], as well as by exploiting time-domain sparsity [77] or through iterative approaches [78, 79]. Alternatively, when the microphone array configuration is known, DOA estimation methods (cf. Sec. 1.1) can also be used in the estimation of the direct-path RTF [63, 80, 81]. Finally, in cases where the spatial information is fully or partially known a priori, alternative MVDR and MWF-based speech enhancement methods have also been proposed [66, 81–85].

1.3 Applications and challenges

The applications of sound source localization and speech enhancement techniques continue to diversify across different domains. In digital telephony, hands-free communication systems and multi-party videoconferencing can benefit from active speaker detection and localization, automatic gain control and noise reduction, such that the robustness and user experience in such applications are improved [6, 60, 63, 86].

Another relevant application of source localization and speech enhancement corresponds to hearing devices. Hearing impairments are known to affect speech intelligibility and spatially selective hearing [87]. Hence, technologies such as hearing aids and cochlear implants greatly benefit from target speaker localization and speech enhancement techniques, improving speech intelligibility and overall listening comfort [8, 75, 82, 88–91].

Regarding human-to-machine interaction, source localization and speech enhancement can improve the performance of automatic speech recognition (ASR) systems [9], enabling digital assistance, voice-based search and other functionalities [92]. The relevance of ASR in improving human-to-machine interaction, and consequently, user experience and accessibility, has led to its integration in smartphones, smart-home systems and in-vehicle technology [9, 93–98].



(a) Multi-party videoconferencing.



(b) Hearing aids.



(c) Smart home systems.

Figure 1.3: Examples of applications of source localization and speech enhancement methods.

Furthermore, considering the potential use of source localization and speech enhancement to detect and classify events in an environment, acoustic monitoring for smart home systems has been developed [11, 99, 100]. Acoustic monitoring has also proven to be relevant for navigation and increased functionality of autonomous robots and vehicles [101–109], as well as of unmanned aerial vehicles (UAVs) [110–112], commonly known as *drones*.

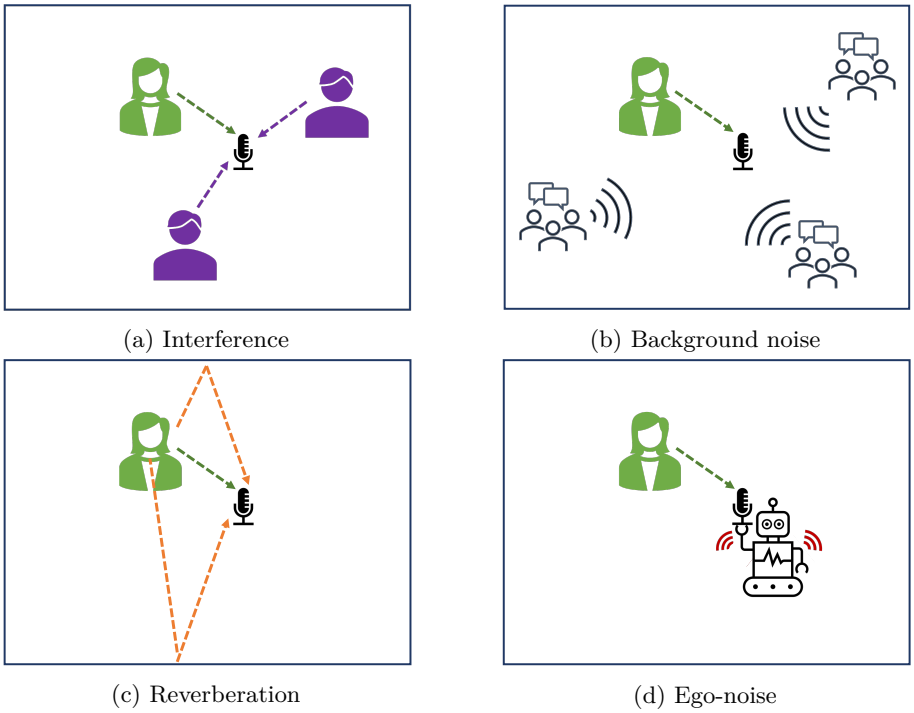
Although the aforementioned applications clearly illustrate the diverse range of opportunities provided by the use of source localization and speech enhancement techniques, the practical conditions in which they are expected to be employed can present many different challenges. We detail the specific challenges addressed in the contributions of this thesis as follows.

1. **Limited setups** - A widely accepted premise in audio processing is that the use of a single microphone limits the possible signal representations to the time and frequency domain [113]. Conversely, a signal representation involving the spatial or wavenumber domain is generally considered to only

be possible for setups with multiple microphones, such that the spatial diversity of the array configuration is considered a fundamental aspect allowing for the localization of sources, as well as the use of spatial filters for speech enhancement [6].

Despite the evident advantages of using microphone arrays, the challenges of exploiting spatial information remain mostly unexplored for devices equipped with a single microphone, as well as for devices in which the fusion of multiple microphone signals is challenging in practice. Examples include mobile phones and smartphones [114], in which simultaneous access to multiple microphones is generally not granted to applications and application developers. In hearing aids, the ongoing miniaturization prevents adequate spacing between microphones [90, 115]. Additionally, wireless acoustic sensor networks encounter challenges such as lack of synchronization, calibration, and network communication between microphones, preventing truly consistent multi-channel signal processing [116].

2. **Adverse acoustic conditions** - Another challenge impacting audio processing for effective source localization and speech enhancement refers to adverse acoustic conditions, encompassing four main aspects in this thesis: interference, background noise, reverberation, and ego-noise. Firstly, interference refers to the signals generated by sources which are considered to be distinctively located in space (and often assumed to be point sources), causing a directional sound field and contributing to the diversity of the observed acoustic environment [6, 60]. Challenges related to interfering sources are often encountered, for instance, in multi-party videoconferencing scenarios, where multiple active speakers are present [6, 60, 63, 86]. Secondly, background noise constitutes unwanted signals arising from a spatially uniform or entirely diffuse sound field, originating from one or multiple sources. Examples of background noise include heating, ventilation, and air conditioning (HVAC) noise, as well as babble noise [6, 60]. Thirdly, reverberation occurs due to the multi-path propagation of sound waves from its source to the receiver, in a room with reflective surfaces [117]. Early reverberation is caused by early reflections, and while it is known to possibly improve speech intelligibility, it also causes coloration in recorded signals [7, 117]. Late reverberation, caused by late reflections, promotes a spatially uniform sound field, similarly to background noise. Although late reverberation plays a role in a listener's spatial perception of the acoustic environment, it tends to degrade speech intelligibility [7, 117]. Lastly, ego-noise specifically refers to the undesired signals generated by the system in which the microphone setup being used is embedded [118]. Ego-noise is observed in applications involving robots, unmanned aerial vehicles (UAVs), or autonomous vehicles, and its



(a) Interference (b) Background noise

(c) Reverberation (d) Ego-noise

Figure 1.4: Examples illustrating the different aspects composing adverse acoustic conditions which pose challenges in source localization and speech enhancement.

characteristics highly depend on the device's structure and functioning [118–120]. Moreover, due to the usual proximity of microphones to the ego-noise sources, its effect over audio processing for source localization and speech enhancement poses significant challenges [118–120]. Illustrations of each of the aspects characterizing adverse acoustic conditions as defined above are presented in Fig. 1.4.

1.4 Contributions

In this section, the contributions to the research problems of source localization and spatial filtering-based speech enhancement presented in this thesis are introduced and associated to the related challenges previously outlined in Sec. 1.3.

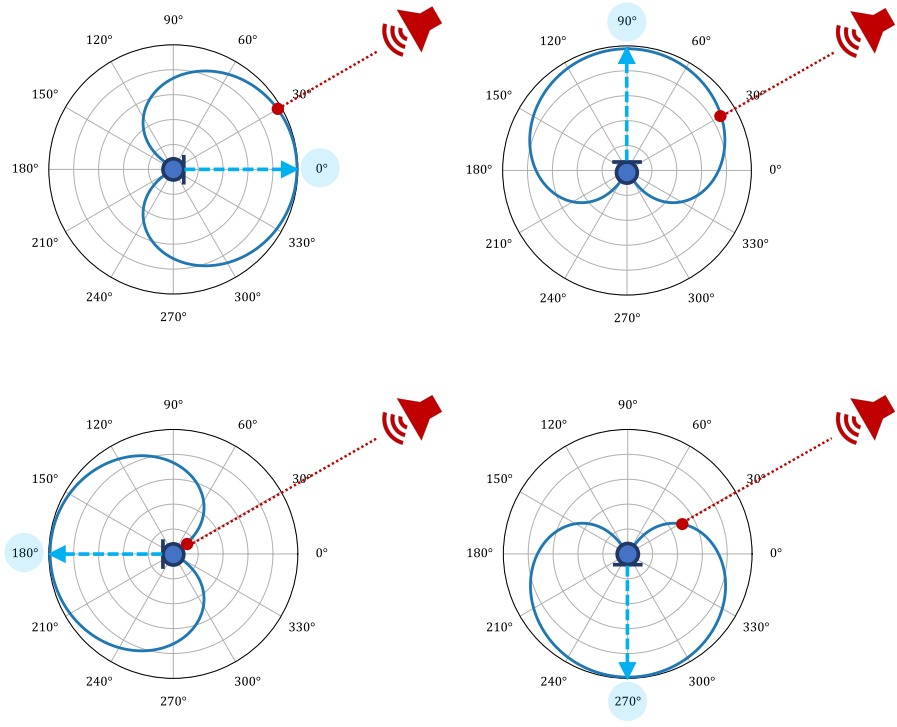


Figure 1.5: Illustration depicting a directional microphone with a cardioid directivity pattern, rotated across various known directions.

1. **Single-channel source localization** - In this thesis, a contribution addressing the challenging task of source localization using a single microphone is presented. Two approaches are proposed for direction-of-arrival (DOA) and power spectral density (PSD) estimation, applicable in scenarios involving single or multiple sound sources.

   The primary focus of this contribution is related to overcoming the limitations posed by a restricted setup that is unfavorable for the estimation of spatial information. The main concept behind the proposed approaches are mainly defined by the following assumptions:

   a. the microphone is directional,

   b. the microphone is allowed to move,

   c. a second sensor in another modality (e.g. accelerometer, camera) is available to infer the microphone orientation, and

d. the sound sources are spatially static point sources.

A directional microphone will capture a spatially static and localized sound source with a different response depending on its orientation, as illustrated in Fig. 1.5. As the microphone orientation varies for different observation time frames, changes in the microphone signal PSD can be analyzed for estimating the DOAs of the sources and their corresponding PSDs [121]. Hence, the rotating setup is identified as an opportunity to facilitate the exploitation of spatial diversity. In both proposed approaches, this is achieved by solving a group-sparse regularized optimization problem for estimating direction-dependent PSD values relative to a given angular dictionary, and locating peaks in the estimated PSD vector.

The first approach, referred to as the Group Lasso Least Squares (GL-LS) approach, is based on the use of a broadband signal model for formulating the group-sparse optimization problem. The DOA estimation is followed by the estimation of the sources' PSDs through the solution of an overdetermined least squares problem. Alternatively, the second approach proposed, which is referred to as the Group Lasso with l_1-reweighting (GL-L1) approach, employs a narrowband signal model when building the linear system of equations used for formulating the group-sparse optimization problem, where the DOAs and PSDs can be jointly estimated by iterative, group-wise reweighting. In both approaches, the number of observations made with the rotating directional microphone can be lower than the total number of candidate DOAs.

The proposed approaches are implemented using the alternating direction method of multipliers (ADMM) [122], and their performance is evaluated through an extensive set of simulations. These simulations involve different setup configurations, ranging from different model mismatches to variations in the acoustic scene and microphone directivity pattern. The results obtained show that the GL-L1 approach can present an advantage over GL-LS in terms of DOA estimation performance in scenarios with low signal-to-noise ratio (SNR) or where multiple sources are closely located to each other. Moreover, it is shown that having the least squares PSD re-estimation step is beneficial in most scenarios tested, such that GL-LS outperformed GL-L1 in terms of PSD estimation errors.

2. **Multi-channel source localization** - In the context of the previous single-channel source localization contribution, one may note that the use of a rotating and directional microphone can be interpreted as the construction of a beamformer in the PSD domain. Accordingly, it can also be interpreted that in the proposed approaches for DOA and PSD estimation, the number of look directions (i.e., microphone orientations) for which the beamformer's output is computed can be lower than the

total number of candidate DOAs considered in the signal model. Building upon this insight, this thesis' second contribution extends the application of the proposed single-microphone source localization framework to multi-microphone setups.

A method is proposed for estimating the DOA of multiple broadband sound sources through the solution of a group-sparse optimization problem, which now models an observed broadband steered response power (SRP) map as a linear combination of PSDs. Based on the sampling requirements of the SRP function, we intend to limit the number of SRP samples relative to the number of candidate DOAs.

The multi-channel DOA estimation method proposed in this study is implemented using the alternating direction method of multipliers (ADMM), and the effect of adverse acoustic conditions over its performance is evaluated through a series of simulations involving multiple speech sources, as well as varying noise and reverberation. In scenarios where the observed SRP map is computed for lower grid resolutions with respect to the spatial sampling requirements of the broadband SRP function, the results obtained show that the proposed method outperforms the conventional SRP-PHAT method. Additionally, it is also shown that the proposed method performs comparably to the group-sparse iterative covariance-based estimation (GSPICE) method. We show that this performance similarity is achieved at a reduced computational cost, indicating the benefit of the proposed approach as an alternative to GSPICE.

3. **Speech enhancement under high levels of ego-noise** - Given the previously outlined contributions on source localization, the final contribution of this thesis aims at exploring the possibilities of effectively employing spatial information in a speech enhancement framework. This is done while considering a practical yet challenging scenario. Specifically, we consider the task of multi-channel speech enhancement using a microphone array embedded in a UAV, and therefore, under considerably adverse acoustic conditions characterized by high levels of ego-noise.

   In order to tackle this challenge, we firstly introduce a brief study on UAV ego-noise, by proposing and demonstrating an experimental methodology for measuring noise emissions from a UAV, operating at constant thrust and simulated hovering condition. Using a hemispherical microphone array setup, the methodology allows for simultaneous noise capture across various elevation and azimuth angles. The methodology is demonstrated using a quadrotor drone, and an analysis on the noise emission levels is performed for a fixed azimuth and varying elevation angles. The results obtained revealed a consistent rise in sound pressure levels (SPLs) across

all observed elevation angles with increasing rotor speed. Additionally, through the analysis of the SPLs in one-third octave bands, it was observed for selected rotor speeds that while the drone noise emission levels may be similar for different elevation angles, some differences in spectral characteristics are present.

The preliminary study on UAV ego-noise is followed by the investigation of potential methods for speech enhancement using a microphone array embedded in a UAV. The incorporation of spatial information is proposed to be performed in two ways. Firstly, we consider exploiting part of the microphone array geometry, in which microphones located near the UAV's propellers are used to capture ego-noise reference signals. The ego-noise references are then employed in the formulation of an MWF-based spatial filtering method, namely the prior knowledge multi-channel Wiener filter (PK-MWF) [83], which exploits prior knowledge on the undesired components for constraining the estimation of the necessary speech correlation matrix. Speech presence probability (SPP) is estimated for detecting speech activity from either an external microphone near the speech source, providing a performance benchmark, or one of the embedded microphones, assuming a more realistic scenario. Secondly, we also consider incorporating spatial information by employing the target speech source's estimated DOA to compute a direct-path RTF, which is then used in an alternative formulation of the PK-MWF.

Experimental measurements are performed in a semi-anechoic chamber, with a UAV mounted on a stand and a loudspeaker playing a speech signal, while considering different average rotor speeds. The recordings obtained and made available online [123] are used to compare the proposed method to the use of the standard multichannel Wiener filter (MWF) estimated with and without the propellers' microphones being used in its formulation, as well as with the use of the direct-path RTF in the alternative formulation of the PK-MWF, based on the DOA estimate obtained when using the multi-channel method presented in this thesis. The results obtained indicate that employing the DOA-based implementation of the PK-MWF filter allows for greater improvement in SNR of the filtered signal, when compared to the result obtained when employing the PK-MWF with SPP estimated from one of the embedded microphones. Conversely, the perceptual improvement, in terms of STOI and PESQ, is compromised. Furthermore, it is shown that when processing different frequency bands independently, as it is commonly done in spatial filtering methods, it is possible to combine the DOA and SPP-based implementations of the PK-MWF. Results indicate that such combination of the two approaches can improve performance in terms of PESQ, at the expense of lower

SNR improvement when compared to the results obtained while solely employing the DOA-based implementation.

1.5 Thesis outline

The remainder of the thesis is organized as follows.

The contributions outlined in Sec. 1.4 are organized in two parts, with **Part I** covering the research problem of source localization, and **Part II** the research problem of speech enhancement.

Part I covers **Ch. 2** and **Ch. 3**. In **Ch. 2**, we present the research work on single-channel source localization. This chapter is based on the publication [124]

- **E. Tengan**, T. Dietzen, F. Elvander, and T. van Waterschoot, "Direction-of-arrival and power spectral density estimation using a single directional microphone and group-sparse optimization," *EURASIP J. Audio, Speech, Music Process.*, vol. 2023, no. 38, p. 1-20, 2023.

In **Ch. 3**, the work on multi-channel source localization is presented. This chapter is based on the submission [125]

- **E. Tengan**, T. Dietzen, F. Elvander and T. van Waterschoot, "Multi-source direction-of-arrival estimation using steered response power and group-sparse optimization", ESAT-STADIUS Tech. Rep. TR 23-163, KU Leuven, Belgium, submitted for publication, Dec. 2023.

Part II covers **Ch. 4**, **Ch. 5** and **Ch. 6**. In **Ch. 4**, the experimental methodology for measuring UAV ego-noise is presented. This chapter is based on the publication [126]

- M. Alkmim, J. Cardenuto, **E. Tengan**, T. Dietzen, T. van Waterschoot, J. Cuenca, L. De Ryck, and W. Desmet, "Drone noise directivity and psychoacoustic evaluation using a hemispherical microphone array," *J. Acoust. Soc. Am.*, vol. 152, no. 5, pp. 2735-2745, 2022.

In **Ch. 5**, the first method proposed for speech enhancement using ego-noise references with a microphone array embedded in a UAV is presented. This chapter is based on the publication [127]

- **E. Tengan**, T. Dietzen, S. Ruiz, M. Alkmim, J. Cardenuto, and T. van Waterschoot, "Speech enhancement using ego-noise references with a microphone array embedded in an unmanned aerial vehicle," in *Proc. 2022 Int. Congr. Acoust. (ICA 2022)*, Gyeongju, South Korea, Oct. 2022, pp. 133-143.

In **Ch. 6**, the proposed incorporation of estimated DOA information in the formulation of the speech enhancement method from **Ch. 5** is presented.

Ch. 7 presents a research valorization plan, where different exploitation points of the contributions presented in this thesis are discussed. Finally, in **Ch. 8**, we conclude the thesis with a summary, as well as with suggestions for future research.

Part I

Source localization

Chapter 2

Single-channel source localization

This chapter is based on the publication:

Direction-of-arrival and power spectral density estimation using a single directional microphone and group-sparse optimization

Elisa Tengan, Thomas Dietzen, Filip Elvander and Toon van Waterschoot

Published in *EURASIP J. Audio, Speech, Music Process.*, vol. 2023, no. 38, p. 1-20, 2023.

Author contributions: All authors jointly developed the methodology and designed the computer simulations presented. ET implemented the algorithms and computer simulations, and all authors jointly interpreted the results obtained. ET drafted the manuscript and all authors read and reviewed the final manuscript.

© 2023 The Authors. Reprinted, with permission, from:
E. Tengan, T. Dietzen, F. Elvander and T. van Waterschoot, “Direction-of-arrival and power spectral density estimation using a single directional microphone and group-sparse optimization," *EURASIP J. Audio, Speech, Music Process.,* vol. 2023, no. 38, p. 1-20, 2023.
Changes include layout, representation and minor editing.

2.1 Introduction

In the field of audio signal processing, the ability to exploit spectral and spatial information from the auditory scene plays an important role in developing speech enhancement, noise reduction and scene analysis techniques [8, 14, 61, 63, 117, 128–130]. The applications in which these tasks have great influence are numerous, with some examples being binaural processing for hearing aids, hands-free telephony and videoconferencing, acoustic surveillance, autonomous robots and so on. By localizing target and interfering sound sources in space and estimating their power spectral densities (PSDs), one can distinguish them and process the recorded signals using the appropriate noise reduction and source classification methods.

Sound source localization is performed by estimating the direction of arrival (DOA) of the signals being recorded. Pioneering methods such as Capon's beamformer [15], the MUSIC algorithm [20] and generalized correlation-based methods [13, 19] are still used for the task of DOA estimation and keep being further modified for improved performance. Alternatively, different approaches have also been developed. Compressed sensing techniques have become more popular [41, 49, 50, 131], with the focus on exploiting sparsity in the signal models considered. Moreover, with the growth of machine learning, data-driven methods have naturally gained more attention as well [132, 133].

In terms of PSD estimation, especially for noise and interfering sources, methods based on minimum statistics [134, 135], minimum mean squared error (MMSE) [136–138] and minima controlled recursive averaging (MCRA) [139–143] have set the baseline for allowing noise reduction and speech enhancement methods to be developed. Later, subspace-based methods [144–146], as well as deep learning methods [147–150] provided different perspectives on how to perform PSD estimation.

Although much has been achieved through the studies presented above, some challenges still remain present. When estimating DOAs, most methods rely on having recordings from multiple microphones, such that their spatial diversity is used for inferring a source's location [6]. However, it is well known that in practice, constraints on hardware design, computational complexity and simultaneous access to multiple microphones' data, often encountered in different devices, may limit the ideal advantages of microphone array processing techniques [62, 151, 152]. In this paper, we propose two alternative approaches for DOA and Power Spectral Density (PSD) estimation using a single, rotating, and directional microphone. By exploring the potential of a single-microphone setup, this study aims to not only provide a single-channel solution that can be more easily adapted to diverse applications, but to also establish a foundational

framework for potential multi-channel extensions that exploit similar principles to those considered in the development of the proposed approaches.

In literature, single-channel DOA estimation methods have already been proposed, with some examples including the use of machine learning [153], a circularly moving microphone that exploits the Doppler effect [154] and a time delay-based subspace approach with a single hydrophone [155]. Our previous work [121] introduced the concept of using a single, rotating directional microphone for performing DOA and PSD estimation. The proposed method involved capturing spatially static and localized sound sources with a cardioid microphone, oriented towards different directions for different observation frames, so that changes in the microphone signal power could be analyzed for determining spatial information about the sources generating the observed sound field. By solving a group-sparsity constrained optimization problem while using a broadband signal model and an overcomplete angular dictionary of possible candidate DOAs for the point sources, DOA estimates could be obtained and used to estimate the localized point sources' PSDs, by solving an overdetermined least squares problem with a nonnegativity constraint for each frequency bin separately. The use of group-sparsity has been previously exploited in multi-channel DOA estimation methods, such as in [156], where a covariance matrix estimated from signals captured at multiple co-prime arrays is modeled based on the corresponding steering vectors and point source PSDs. Moreover, estimating PSDs through the solution of a least squares problem has been also performed in [157], in which multiple beamformers are applied to microphone array signals and their outputs are used as the proposed method's observed data. Nonetheless, when employing the mentioned multi-channel DOA and PSD estimation methods, it becomes necessary to assess the performance constraints associated with the microphone array configuration available in the considered scenario. Factors that could potentially influence these limitations include the spacing between microphones and the overall array geometry [6,158].

In this paper, the method proposed in [121] is extended in two aspects, resulting in two alternative approaches. First, in order to add robustness in scenarios with a mismatch between the signal model and the generated data, we propose to use an additional nonnegativity constraint on the residual noise term when solving the group-sparse optimization problem. We refer to this as the *Group-Lasso Least Squares* (GL-LS) approach. Note that GL-LS is still based on the broadband signal model presented in [121]. Second, in addition to the new non-negativity constraint, we also propose an optimization problem based on a narrowband signal model, where the DOAs and PSDs can be jointly estimated by iterative, group-wise reweighting. This is referred to as the *Group-Lasso with l_1-reweighting* (GL-L1) approach. An efficient implementation of both approaches using the alternating direction method of multipliers (ADMM) is

presented, as opposed to the use of CVX [159] as it was done in our previous work [121].

In order to evaluate the performance of the proposed approaches, a series of simulations subject to different types of model mismatch are performed, by introducing off-grid DOAs, non ideal microphone responses and reverberation for the case of a single point source. We also analyze scenarios involving two point sources of different broadband power and varying angular separation between them. Finally, we consider the case of three point sources when using higher-order directivity patterns. These simulations greatly extend the aspects considered for evaluating the proposed approaches in comparison with those presented in our previous work [121]. We show that most of the times, having the PSD re-estimation step with least squares as performed with GL-LS is beneficial, but also that redesigning the system of equations in a frequency-dependent manner as performed in GL-L1 may help in distinguishing closely located sources.

This paper is structured as follows. In Sec. 2.2, we present the signal model. In Sec. 2.3, we explain the proposed approaches. In Sec. 2.4, we explain the ADMM-based implementation. In Sec. 2.5, we present the simulations setup, the results obtained and discussion. Finally, in section 2.6, we conclude with a summary of the work presented and future work.

2.2 Signal model

2.2.1 Rotational microphone signal model

In the short-time Fourier transform (STFT) domain, the signal recorded by a single, directional microphone rotating in the horizontal plane is modeled as

$$\begin{aligned} Y(k,n) = &\sum_{p=1}^{P} S_p(k,n) \int_0^{2\pi} a^*(k,\theta-\gamma_n) H_p(k,\theta) d\theta \\ &+ D(k,n), \end{aligned} \tag{2.1}$$

where $(\cdot)^*$ denotes the complex conjugate, k the discrete frequency index, n the observation frame index, θ the look direction and γ_n the microphone orientation at frame n. We assume that there are a total of P point sources in the far field and that P is known. The expression in (2.1) describes that the resulting microphone signal $Y(k,n)$ is composed of the sum of the P point

source signals $S_p(k,n)$ arriving from P distinct directions, ϑ_1 to ϑ_P, weighted by the direction-dependent microphone response $a(k,\theta-\gamma_n)$, relative to the microphone orientation γ_n, and by the room transfer function (RTF) $H_p(k,\theta)$ from the p-th far-field source to the microphone, added to diffuse or sensor noise $D(k,n)$. If we consider that the recording is performed in anechoic conditions, then $H_p(k,\theta)=\delta(\theta-\vartheta_p)$, $\forall\, k$ and for $p=1,\ldots,P$, such that the expression in (2.1) is reduced to

$$Y(k,n)=\sum_{p=1}^{P} a^*(k,\vartheta_p-\gamma_n)S_p(k,n)+D(k,n). \tag{2.2}$$

Assuming that the source signals are uncorrelated and stationary during the entire observation, such that their PSDs remain constant across different time frames, and that the microphone response is real-valued, the microphone signal PSD $\phi_Y(k,n)$ can be described as follows:

$$\phi_Y(k,n)=\sum_{p=1}^{P} |a(k,\vartheta_p-\gamma_n)|^2\phi_{S_p}(k)+\phi_D(k,n), \tag{2.3}$$

where $\phi_D(k,n)$ is the noise PSD for frequency k and time frame n, and $\phi_{S_p}(k)$ is the PSD for frequency k corresponding to the p-th source at position ϑ_p. As the directional microphone is oriented towards different directions γ_n for different observation frames n, the relative positions of the sound sources with respect to the microphone do not remain the same, and consequently their PSD values $\phi_{S_p}(k)$ are multiplied with different squared microphone response coefficients $|a(k,\vartheta_p-\gamma_n)|^2$ over different time frames.

2.2.2 Grid-based system of equations

If we assume that measurements of $\phi_Y(k,n)$ are available for multiple time frames, with $n=1,\ldots,N$, a linear system of equations can be built for estimating the point source DOAs ϑ_p and PSDs $\phi_{S_p}(k)$, for $p=1,\ldots,P$ and $k=1,\ldots,K$, where K denotes the number of frequency bins. As the directional weighting factor $|a(k,\vartheta_p-\gamma_n)|^2$ in (2.3) depends on the unknown source DOA, we use a grid of candidate positions defined between 0 and 2π and an overcomplete dictionary of corresponding microphone response coefficients in order to build the linear system of equations, as previously proposed in [121].

For a single candidate angle, denoted as θ_l, a vector $\boldsymbol{\varphi}_{S,\theta_l}$ with PSD values for all frequencies is defined as

$$\boldsymbol{\varphi}_{S,\theta_l}=\left[\phi_S(1,\theta_l)\ \ \ldots\ \ \phi_S(K,\theta_l)\right]^\top, \tag{2.4}$$

where $(\cdot)^\top$ denotes the transpose and $\phi_S(k, \theta_l)$ is the PSD at candidate position θ_l for frequency k. We stack different $\boldsymbol{\varphi}_{S,\theta_l}$ vectors for all candidate DOAs from a given L-element angular grid, with $P \ll L$ and $N < L$, in order to obtain the vector we ultimately aim to estimate as

$$\boldsymbol{\varphi}_S = \left[\boldsymbol{\varphi}_{S,\theta_1}^\top \ \ldots \ \boldsymbol{\varphi}_{S,\theta_L}^\top\right]^\top, \tag{2.5}$$

which will be used to obtain $\hat{\vartheta}_p$ and $\hat{\phi}_{S_p}(k)$ such that $\hat{\vartheta}_p \in \{\theta_1, \ldots, \theta_L\}$, for $p = 1, \ldots, P$.

One possible construction of the linear system of equations is defined as follows. Firstly, a vector $\boldsymbol{\varphi}_{Y,n}$ containing the PSD values for the microphone observation frame n is defined as

$$\boldsymbol{\varphi}_{Y,n} = \left[\phi_Y(1, n) \ \ldots \ \phi_Y(K, n)\right]^\top. \tag{2.6}$$

By stacking different $\boldsymbol{\varphi}_{Y,n}$ vectors for all observations $n = 1$ to $n = N$, we obtain a vector $\boldsymbol{\varphi}_Y$ as

$$\boldsymbol{\varphi}_Y = \left[\boldsymbol{\varphi}_{Y,1}^\top \ \ldots \ \boldsymbol{\varphi}_{Y,N}^\top\right]^\top. \tag{2.7}$$

Similarly, we also define the vectors for the diffuse component as

$$\boldsymbol{\varphi}_{D,n} = \left[\phi_D(1, n) \ \ldots \ \phi_D(K, n)\right]^\top, \tag{2.8}$$

$$\boldsymbol{\varphi}_D = \left[\boldsymbol{\varphi}_{D,1}^\top \ \ldots \ \boldsymbol{\varphi}_{D,N}^\top\right]^\top. \tag{2.9}$$

In order to build an overcomplete microphone response matrix involving all possible candidate DOAs, we firstly define a vector containing the squared, microphone response for a candidate angle θ_l relative to the microphone orientation γ_n as

$$\mathbf{a}(\theta_l - \gamma_n) = \left[|a(1, \theta_l - \gamma_n)|^2 \ \ldots \ |a(K, \theta_l - \gamma_n)|^2\right]. \tag{2.10}$$

Hence, the microphone response matrix is defined as

$$\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A}_{1,1} & \ldots & \mathbf{A}_{1,L} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{N,1} & \ldots & \mathbf{A}_{N,L} \end{bmatrix}, \tag{2.11}$$

where $\mathbf{A}_{n,l} = \text{diag}\left(\mathbf{a}(\theta_l - \gamma_n)\right)$ and $\text{diag}(\cdot)$ denotes the creation of a diagonal matrix from the elements of a given vector. The linear system of equations can then be written as

$$\boldsymbol{\varphi}_Y = \bar{\mathbf{A}}\boldsymbol{\varphi}_S + \boldsymbol{\varphi}_D, \tag{2.12}$$

which corresponds to a matrix representation of (2.3) for different observation frames, and which we refer to as the narrowband system of equations in the remainder.

By ensuring that $\gamma_1 \neq \gamma_2 \neq \ldots \neq \gamma_N$, with $0 \leq \gamma_n \leq 2\pi$, $\forall\, n$, and assuming that $\boldsymbol{\varphi}_Y$ and $\bar{\mathbf{A}}$ are known, source localization in terms of DOA estimation can be achieved by solving the proposed linear system of equations. From the estimated vector $\hat{\boldsymbol{\varphi}}_S$ we can identify in which direction θ_l within the angular grid there are peaks in power, indicating the point source DOAs $\hat{\vartheta}_p$ and their PSDs $\hat{\phi}_{S_p}(k)$, assuming that $\hat{\vartheta}_p \in \{\theta_1, \ldots, \theta_L\}$.

While this narrowband system of equations has all frequency bins decoupled from each other, it is also possible to build a broadband version of it. By considering an $N \times N$ identity matrix denoted as $\mathbf{I}_N$ and a K-element column vector of ones denoted as $\mathbf{1}_K$, we define the broadband vectors and matrix:

$$\widetilde{\boldsymbol{\varphi}}_Y = \left(\mathbf{I}_N \otimes \mathbf{1}_K^\top\right) \boldsymbol{\varphi}_Y, \tag{2.13}$$

$$\widetilde{\boldsymbol{\varphi}}_D = \left(\mathbf{I}_N \otimes \mathbf{1}_K^\top\right) \boldsymbol{\varphi}_D, \tag{2.14}$$

$$\widetilde{\mathbf{A}} = \left(\mathbf{I}_N \otimes \mathbf{1}_K^\top\right) \bar{\mathbf{A}}, \tag{2.15}$$

where $\otimes$ denotes the Kronecker product. This operation over the previously defined vectors $\boldsymbol{\varphi}_Y$ and $\boldsymbol{\varphi}_D$ results in the sum over all frequency bins of the PSD values from the microphone signal and the diffuse component, respectively. It also results in a 'dimensionality reduction' of the original matrix $\bar{\mathbf{A}}$ such that $\mathbf{A}_{n,l}$ is replaced by $\mathbf{a}(\theta_l - \gamma_n)$ stacked together accordingly. Using (2.13)-(2.15), the linear system of equations can be written as

$$\widetilde{\boldsymbol{\varphi}}_Y = \widetilde{\mathbf{A}} \boldsymbol{\varphi}_S + \widetilde{\boldsymbol{\varphi}}_D, \tag{2.16}$$

which we refer to as the broadband system of equations in the remainder, which is equivalent to the system of equations used in [121]. Like the narrowband system presented in (2.12), the broadband system in (2.16) can be solved for estimating the point source DOAs, with the caveat of having lost the frequency-dependent information in the microphone observation vector by summing the target input PSD vector over all frequencies. Consequently, an additional PSD re-estimation step is required when using the broadband system of equations as will be explained further in the following section.

2.3 Proposed approaches

In this section, we explain the broadband and narrowband approach proposed for estimating the DOAs and PSDs of point sources modeled as previously

described in Sec. 2.2. The first approach involves the solution to a group-sparse constrained optimization problem for estimating the DOAs, constructed based on the broadband system of equations in (2.16), followed by a least-squares step for re-estimating the point sources' PSDs. This approach is here named GL-LS. The second approach, involving the solution to a group-sparse constrained optimization problem based on the narrowband system of equations in (2.12), considers an iterative process of reweighting the group-sparsity penalty present in the formulation for jointly estimating the DOAs and PSDs of the point sources. This approach is here named GL-L1.

2.3.1 Group Lasso followed by least squares (GL-LS)

Assuming that $N < KL$, the linear system in (2.16) is underdetermined. Therefore, the following Group Lasso [160] optimization problem is proposed to be solved:

$$\underset{\boldsymbol{\varphi}_S}{\text{minimize}} \quad \frac{1}{2}\left\|\widetilde{\boldsymbol{\varphi}}_Y - \widetilde{\mathbf{A}}\boldsymbol{\varphi}_S\right\|_2^2 + \lambda \sum_{l=1}^{L} \|\boldsymbol{\varphi}_{S,\theta_l}\|_2 \tag{2.17a}$$

$$\text{subject to} \quad \boldsymbol{\varphi}_S \geq 0 \tag{2.17b}$$

$$\widetilde{\boldsymbol{\varphi}}_Y - \widetilde{\mathbf{A}}\boldsymbol{\varphi}_S \geq 0 \tag{2.17c}$$

The nonnegativity constraint in (2.17b) is necessary for complying with the intrinsic nonnegativity property of PSD values [161], whereas the nonnegativity constraint over the error term, as formulated in (2.17c), is included as a means to model the noise signal $\phi_D(k,n)$ present in the microphone observations, such that more robustness can be achieved in case of a possible model mismatch, for instance between the assumed matrix $\widetilde{\mathbf{A}}$ and the actual microphone response. The Group Lasso formulation includes the regularization term $\lambda \sum_{l=1}^{L} \|\boldsymbol{\varphi}_{S,\theta_l}\|_2$, which enforces sparsity between so-called different groups [160]. When assuming that only a limited number of point sources are present in space ($P \ll L$), using this group-sparsity penalty is a way of ensuring that only a few of the subvectors $\boldsymbol{\varphi}_{S,\theta_l}$ composing $\boldsymbol{\varphi}_S$ will be activated, that is, have a magnitude significantly different from zero. After solving the optimization problem (2.17a)-(2.17c) and obtaining $\hat{\boldsymbol{\varphi}}_S$, and therefore, $\hat{\boldsymbol{\varphi}}_{S,\theta_1}, \ldots, \hat{\boldsymbol{\varphi}}_{S,\theta_L}$, the PSD values can be averaged over the K frequency bins for each of the L candidate directions, allowing the point source DOAs to be estimated by finding the indices of θ_l for which there are peaks in the average PSD. For a total of P sources assumed to be present, P peaks should then be identified.

One may note that the resulting PSD estimates for all directions in the angular grid obtained from solving (2.17a)-(2.17c) will be inherently biased, due to the group-sparsity penalty included in the optimization problem [162]. Moreover, the summation over frequency of the PSD values present in $\widetilde{\boldsymbol{\varphi}}_Y$, see (2.13), results in the loss of frequency-dependent information on the detected point sources' PSDs. These limiting factors motivate the use of a re-estimation step for the PSD values using the estimated DOAs, as previously proposed in [121].

The new PSD vectors are defined as follows:

$$\boldsymbol{\phi}_Y(k) = \left[\phi_Y(k,1) \;\; \dots \;\; \phi_Y(k,N)\right]^\top, \tag{2.18}$$

$$\boldsymbol{\phi}_S(k) = \left[\phi_{S_1}(k) \;\; \dots \;\; \phi_{S_P}(k)\right]^\top, \tag{2.19}$$

$$\boldsymbol{\phi}_D(k) = \left[\phi_D(k,1) \;\; \dots \;\; \phi_D(k,N)\right]^\top. \tag{2.20}$$

We define a new matrix $\mathbf{A}_S(k) \in \mathbb{R}^{N\times P}$ which now contains squared microphone response coefficients for only the directions $\hat{\vartheta}_1, \dots, \hat{\vartheta}_P \in \{\theta_1, \dots, \theta_L\}$ where the P sources are assumed to be located, based on the preceding DOA estimation:

$$\mathbf{A}_S(k) = \begin{bmatrix} |a(k,\hat{\vartheta}_1 - \gamma_1)|^2 & \dots & |a(k,\hat{\vartheta}_P - \gamma_1)|^2 \\ \vdots & & \vdots \\ |a(k,\hat{\vartheta}_1 - \gamma_N)|^2 & \dots & |a(k,\hat{\vartheta}_P - \gamma_N)|^2 \end{bmatrix}. \tag{2.21}$$

Using the PSD signal model in (2.3), a new linear system of equations for the microphone signal PSD is then formulated, for each frequency bin, as

$$\boldsymbol{\phi}_Y(k) = \mathbf{A}_S(k)\boldsymbol{\phi}_S(k) + \boldsymbol{\phi}_D(k). \tag{2.22}$$

If $P \leq N$, a constrained least-squares approach can be used for solving the overdetermined linear system and estimating the PSD values of the point sources:

$$\underset{\boldsymbol{\phi}_S(k)}{\text{minimize}} \quad \frac{1}{2}\left\|\boldsymbol{\phi}_Y(k) - \mathbf{A}_S(k)\boldsymbol{\phi}_S(k)\right\|_2^2 \tag{2.23a}$$

$$\text{subject to} \quad \boldsymbol{\phi}_S(k) \geq 0 \tag{2.23b}$$

Hence, in this re-estimation step, we avoid the bias induced by the Group Lasso formulation presented in the DOA estimation step and allow for a more accurate PSD estimation for the stationary point sources.

When compared to our previous work [121], the GL-LS approach described in this subsection presents an extension of the previous idea of solving a group-sparse constrained optimization problem for estimating DOAs and PSDs of point sources by including the additional nonnegativity constraint over the error term, in order to provide robustness against model mismatches that may be present.

2.3.2 Group Lasso with l_1-reweighting (GL-L1)

As an alternative to the method proposed in Sec. 2.3.1, we consider employing the narrowband input vector $\boldsymbol{\varphi}_Y$ and response matrix $\bar{\mathbf{A}}$ (see the model in (2.12)), and construct the following optimization problem:

$$\underset{\boldsymbol{\varphi}_S}{\text{minimize}} \quad \frac{1}{2}\left\|\boldsymbol{\varphi}_Y - \bar{\mathbf{A}}\boldsymbol{\varphi}_S\right\|_2^2 + \lambda \sum_{l=1}^{L} w_l \left\|\boldsymbol{\varphi}_{S,\theta_l}\right\|_2 \tag{2.24a}$$

$$\text{subject to} \quad \boldsymbol{\varphi}_S \geq 0 \tag{2.24b}$$

$$\boldsymbol{\varphi}_Y - \bar{\mathbf{A}}\boldsymbol{\varphi}_S \geq 0 \tag{2.24c}$$

In this approach, we include the use of group weights, denoted w_l for $l = 1, \ldots, L$, in order to allow for a group-wise, iterative reweighting process aimed at enhancing sparsity within the final solution obtained [163, 164]. This is motivated by the goal of jointly estimating the point source DOAs and PSDs from the solution to the constrained optimization problem, without performing a least-squares re-estimation step as in the GL-LS approach, in which the optimization problem in (2.17) is solved only for estimating the source's DOAs.

The group sparsity is enhanced by successively solving the optimization problem in (2.24), while updating, between iterations, the sparsity penalization of each estimated group separately as a function of their corresponding norms [163, 165, 166]. The weight updates can be expressed as

$$w_l^{(i+1)} = 1/(\|\hat{\boldsymbol{\varphi}}_{S,\theta_l}^{(i)}\|_2 + \epsilon) \quad \text{for} \quad l = 1, \ldots, L, \tag{2.25}$$

where i denotes the reweighting iteration index, and $\epsilon > 0$ is a fixed parameter introduced for avoiding a division by zero.

After reaching convergence from the group-wise iterative reweighting procedure, i.e., repeatedly re-estimating $\hat{\boldsymbol{\varphi}}_{S,\theta_l}$ while updating the sparsity penalty weights

for different groups individually, the point source DOA estimation is performed in a similar fashion as proposed in Sec. 2.3.1. Assuming a total number of P point sources, the P largest peaks in frequency-averaged power obtained from $\hat{\boldsymbol{\varphi}}_S$ are picked, and the estimated DOAs $\hat{\vartheta}_p$ are obtained from the indices of θ_l for the corresponding groups selected.

Simultaneously, the point source PSDs are obtained as the estimated values of the corresponding subvectors contained in $\hat{\boldsymbol{\varphi}}_S$ with angles $\hat{\vartheta}_p$, for $p = 1, \ldots, P$. Hence, in this approach, the PSD estimates are obtained jointly with the DOA estimates, without performing an additional least-squares re-estimation step as in the proposed GL-LS approach.

The GL-L1 approach described in this subsection presents an extension of our previous work [121] in more aspects than GL-LS, namely the employment of the additional nonnegativity constraint over the error term, the use of the narrowband linear system of equations and the iterative group reweighting process for jointly estimating DOAs and PSDs without a least-squares re-estimation step.

2.4 Implementation

In order to solve the optimization problems defined in (2.17) and (2.24) in an efficient way, we employ the alternating direction method of multipliers (ADMM) algorithm [122], as opposed to the use of CVX [159] as it was done in our previous work [121]. Since in both proposed approaches the vector $\boldsymbol{\varphi}_S$ to be estimated is the same, and their distinction only consists in the construction of the input vectors and matrices employed, as well as in the use of group weights, we firstly describe the used ADMM implementation in general terms and then further clarify each method's full algorithm afterwards.

Let a general optimization problem involving the same group sparsity and nonnegativity constraints present in (2.17) and (2.24) be described as

$$\underset{\mathbf{x}}{\text{minimize}} \quad \frac{1}{2}\left\|\mathbf{b}-\mathbf{A}\mathbf{x}\right\|_2^2 + \lambda \sum_{l=1}^{L} w_l \left\|\mathbf{x}_l\right\|_2 \tag{2.26a}$$

$$\text{subject to} \quad \mathbf{x} \geq 0 \tag{2.26b}$$

$$\mathbf{b}-\mathbf{A}\mathbf{x} \geq 0 \tag{2.26c}$$

where $\mathbf{x}_l$ corresponds to the l-th group composing the vector $\mathbf{x}$. We can use auxiliary variables, denoted $\mathbf{u}_1$ and $\mathbf{u}_2$, to rewrite the problem as

$$\underset{\mathbf{u}_1 \geq 0,\ \mathbf{u}_2 \geq 0}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{u}_1\|_2^2 + \lambda \sum_{l=1}^{L} w_l \|\mathbf{u}_{2,l}\|_2 \tag{2.27a}$$

$$\text{subject to} \quad \mathbf{u}_1 = \mathbf{b} - \mathbf{A}\mathbf{x} \tag{2.27b}$$

$$\mathbf{u}_2 = \mathbf{x} \tag{2.27c}$$

where $\mathbf{u}_{2,l}$ corresponds to the l-th group composing the vector $\mathbf{u}_2$, such that it can be represented as

$$\mathbf{u}_2 = \begin{bmatrix} \mathbf{u}_{2,1}^\top & \dots & \mathbf{u}_{2,L}^\top \end{bmatrix}^\top. \tag{2.28}$$

The augmented Lagrangian [122] of the optimization problem in (2.27) can be written as

$$\begin{aligned} \mathcal{L}_\rho(\mathbf{x}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{d}_1, \mathbf{d}_2) = {} & \frac{1}{2} \|\mathbf{u}_1\|_2^2 + \lambda \sum_{l=1}^{L} w_l \|\mathbf{u}_{2,l}\|_2 \\ & + \frac{\rho}{2} \|\mathbf{b} - \mathbf{A}\mathbf{x} - \mathbf{u}_1 - \mathbf{d}_1\|_2^2 \\ & + \frac{\rho}{2} \|\mathbf{x} - \mathbf{u}_2 - \mathbf{d}_2\|_2^2 \end{aligned} \tag{2.29}$$

where $\mathbf{d}_1$ and $\mathbf{d}_2$ are the dual variables, and ρ can be interpreted as a dual update step size [122].

Considering an ADMM iteration indexed as j, we define the following short hands:

$$\boldsymbol{\zeta}_1(j) = \mathbf{b} - \mathbf{A}\mathbf{x}(j+1) - \mathbf{d}_1(j), \tag{2.30}$$

$$\boldsymbol{\zeta}_2(j) = \mathbf{x}(j+1) - \mathbf{d}_2(j). \tag{2.31}$$

The updates for each variable in (2.29) are

$$\mathbf{x}(j+1) = \arg\min_{\mathbf{x}} \mathcal{L}_\rho(\mathbf{x}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{d}_1, \mathbf{d}_2), \tag{2.32}$$

$$\mathbf{u}_1(j+1) = \arg\min_{\mathbf{u}_1 \geq 0} \frac{1}{2} \|\mathbf{u}_1\|_2^2 + \frac{\rho}{2} \|\boldsymbol{\zeta}_1(j) - \mathbf{u}_1\|_2^2, \tag{2.33}$$

$$\mathbf{u}_2(j+1) = \arg\min_{\mathbf{u}_2 \geq 0} \lambda \sum_{l=1}^{L} w_l \|\mathbf{u}_{2,l}\|_2 + \frac{\rho}{2} \|\boldsymbol{\zeta}_2(j) - \mathbf{u}_2\|_2^2, \tag{2.34}$$

$$\mathbf{d}_1(j+1) = \mathbf{u}_1(j+1) - \boldsymbol{\zeta}_1(j), \tag{2.35}$$

$$\mathbf{d}_2(j+1) = \mathbf{u}_2(j+1) - \boldsymbol{\zeta}_2(j). \tag{2.36}$$

The first update in (2.32) is computed as

$$\begin{aligned} \mathbf{x}\big(j+1\big) = \big(\mathbf{A}^\top \mathbf{A} + \mathbf{I}\big)^{-1} \big[\mathbf{A}^\top (\mathbf{b} - \mathbf{u}_1(j) - \mathbf{d}_1(j)) \\ + \mathbf{u}_2(j) + \mathbf{d}_2(j)\big]. \end{aligned} \tag{2.37}$$

The $\mathbf{u}_1$ update in (2.33) takes into account its nonnegativity constraint and is computed as

$$\mathbf{u}_1\big(j+1\big) = \max\left(0, \frac{\rho \boldsymbol{\zeta}_1(j)}{1+\rho}\right), \tag{2.38}$$

where $\max(\cdot)$ denotes the element-wise max operator.

The update of $\mathbf{u}_2$ in (2.34) is computed as

$$\mathbf{u}_{2,l}\big(j+1\big) = \mathrm{T}\left(\max\left(\boldsymbol{\zeta}_{2,l}(j), 0\right), w_l \frac{\lambda}{\rho}\right), \tag{2.39}$$

for $l = 1, \ldots, L$, where $\mathrm{T}(\cdot)$ represents a group-wise shrinkage function, defined as

$$\mathrm{T}(\mathbf{z}, \kappa) = \left[\max\left(1 - \frac{\kappa}{\|\mathbf{z}\|_2}, 0\right)\right] \mathbf{z}. \tag{2.40}$$

After computing the updates of $\mathbf{d}_1$ and $\mathbf{d}_2$ as in (2.35) and (2.36), respectively, the whole iterative process is repeated until convergence [122]. For more detailed derivations of each ADMM update equation, we refer to [122].

Algorithm 1: The proposed GL-LS algorithm

initiate $\mathbf{A} = \widetilde{\mathbf{A}}$, $\mathbf{b} = \widetilde{\boldsymbol{\varphi}}_Y$
  $\hat{\boldsymbol{\varphi}}_S = \mathbf{0}^{KL\times 1}$
  $w_l = 1$ for $l = 1, \ldots, L$
initiate $j := 0$, $\mathbf{u}_2(0) = \hat{\boldsymbol{\varphi}}_S$,
  $\mathbf{x}(0) = \mathbf{d}_1(0) = \mathbf{d}_2(0) = \mathbf{0}^{KL\times 1}$
repeat {ADMM scheme}
  update $\mathbf{x}(j+1)$ as in (2.37)
  update $\mathbf{u}_1\big(j+1\big)$ as in (2.38)
  for $l \leftarrow 1$ to L **do**
    update $\mathbf{u}_{2,l}$ as in (2.39)
  end for
  $\mathbf{d}_1(j+1) = \mathbf{u}_1(j+1) - \boldsymbol{\zeta}_1(j)$
  $\mathbf{d}_2(j+1) = \mathbf{u}_2(j+1) - \boldsymbol{\zeta}_2(j)$
  $j \leftarrow j+1$
until convergence
store $\hat{\boldsymbol{\varphi}}_S = \mathbf{u}_2(\text{end})$
estimate $\hat{\vartheta}_p$ for $p = 1, \ldots, P$ as the P largest peaks averaged over frequency in $\hat{\boldsymbol{\varphi}}_S$
estimate $\hat{\boldsymbol{\phi}}_S(k)$ in (2.23) for $k = 1, \ldots, K$

As previously mentioned, the difference in implementation between the proposed methods GL-LS and GL-L1 relies on how the input data structure for the ADMM algorithm here explained is chosen. In the case of GL-LS, we use $\widetilde{\boldsymbol{\varphi}}_Y$ and $\widetilde{\mathbf{A}}$ as input vector and response matrix, respectively, and we set all group weights w_l equal to one. After running the ADMM scheme once, the result is used for estimating the source DOAs and re-estimating the source PSDs as explained in Sec. 2.3.1. In the case of GL-L1, we use $\boldsymbol{\varphi}_Y$ and $\bar{\mathbf{A}}$ as input, the weights are also initially set equal to one, and the ADMM scheme is repeated until convergence while re-updating the group weights as described in Sec. 2.3.2, so that a final joint DOA and PSD estimation is obtained. A summary of the GL-LS implementation is presented in Algorithm 1, while a summary of the GL-L1 implementation is presented in Algorithm 2.

Regarding the computational complexity of each proposed approach, numerous factors will affect the overall cost of both GL-LS and GL-L1. Firstly, when analysing a single iteration within the ADMM scheme, we can observe that the most computationally demanding update occurs in (2.37), and that its cost will depend on the dimensions of the response matrix being employed [122]. The use of the wideband signal model by GL-LS and the narrowband signal model by GL-L1 results in an assymptotic complexity of $\mathcal{O}(NKL)$ and $\mathcal{O}(NK^2L)$,

Algorithm 2: The proposed GL-L1 algorithm

initiate $i := 0$, $\mathbf{A} = \bar{\mathbf{A}}$, $\mathbf{b} = \boldsymbol{\varphi}_Y$,
  $\hat{\boldsymbol{\varphi}}_S^{(0)} = \mathbf{x}_{\text{save}} = \mathbf{d}_{1,\text{save}} = \mathbf{d}_{2,\text{save}} = \mathbf{0}^{KL\times 1}$,
  $w_l^{(0)} = 1$ for $l = 1, \ldots, L$
repeat {reweighting scheme}
  initiate $j := 0$, $\mathbf{u}_2(0) = \hat{\boldsymbol{\varphi}}_S^{(i)}$,
    $\mathbf{x}(0) = \mathbf{x}_{\text{save}}$, $\mathbf{d}_1(0) = \mathbf{d}_{1,\text{save}}$, $\mathbf{d}_2(0) = \mathbf{d}_{2,\text{save}}$
  repeat {ADMM scheme}
    update $\mathbf{x}(j+1)$ as in (2.37)
    update $\mathbf{u}_1\big(j+1\big)$ as in (2.38)
    for $l \leftarrow 1$ to L **do**
      update $\mathbf{u}_{2,l}$ as in (2.39)
    end for
    $\mathbf{d}_1(j+1) = \mathbf{u}_1(j+1) - \boldsymbol{\zeta}_1(j)$
    $\mathbf{d}_2(j+1) = \mathbf{u}_2(j+1) - \boldsymbol{\zeta}_2(j)$
    $j \leftarrow j+1$
  until convergence
  store $\hat{\boldsymbol{\varphi}}_S^{(i)} = \mathbf{u}_2(\text{end})$, $\mathbf{x}_{\text{save}} = \mathbf{x}(\text{end})$, $\mathbf{d}_{1,\text{save}} = \mathbf{d}_1(\text{end})$,
  $\mathbf{d}_{2,\text{save}} = \mathbf{d}_2(\text{end})$
  update $w_l^{(i+1)} = 1/(\|\hat{\boldsymbol{\varphi}}_{S,\theta_l}^{(i)}\|_2 + \epsilon)$ for $l = 1, \ldots, L$
  $i \leftarrow i+1$
until convergence
estimate $\hat{\vartheta}_p$ for $p = 1, \ldots, P$ as the P largest peaks averaged over frequency in $\hat{\boldsymbol{\varphi}}_S$
store $\hat{\boldsymbol{\phi}}_S(k)$ based on $\hat{\boldsymbol{\varphi}}_S$ and $\hat{\vartheta}_p$

respectively, for each of the proposed approaches. In the case of GL-LS, the cost of the additional PSD re-estimation step through the solution of a least squares problem will assymptotically be $\mathcal{O}(N^3)$. Finally, the total runtime experienced by each proposed approach will depend on the convergence criterion selected for the ADMM scheme and the number of iterations required for satisfying it. Therefore, even though the use of the proposed approaches may significantly differ in overall computational cost depending on the number of microphone orientations N being considered and setup conditions that can affect the convergence rate, the GL-LS approach is currently more computationally efficient than the GL-L1 approach. One may note, however, that the sparse structure of the matrix employed in the narrowband signal model, see (2.11), allows for using sparsity-aware methods for matrix computations [167, 168] that can potentially reduce the complexity of GL-L1.

2.5 Simulations

In order to evaluate the performance of both proposed approaches in terms of DOA and PSD estimation, simulations are done in MATLAB considering different setups. We aim to observe the advantages or disadvantages of using GL-LS or GL-L1 under different conditions, ranging from different types of model mismatch in Sec. 2.5.2-2.5.3 to variations of the acoustic scene in Sec. 2.5.4-2.5.5 and of the microphone directivity pattern in Sec. 2.5.6.

In Table 2.1, a summary of the main parameters used in each subsection is presented. For all simulations, the sampling frequency is 16 kHz, the source signals are stationary and the microphone signal PSD $\phi_Y(k,n)$ is estimated with Welch's method, using a 512-point Hann window corresponding to a length of 32 ms and 50% overlap over a duration of 500 ms for each microphone orientation γ_n, and therefore, each observation frame n. For a total of N observations, N different microphone orientations uniformly distributed over 360° are simulated. The L candidate directions used for building the microphone response matrices are defined according to a uniformly spaced angular grid given a certain resolution in degrees. The signal-to-noise ratio (SNR) is defined as $\text{SNR} = \sigma_{S_1}^2/\sigma_D^2$, where $\sigma_{S_1}^2$ denotes the broadband power of the first source ($p = 1$), and σ_D^2 denotes the broadband power of the diffuse component. The regularization parameter λ is heuristically set as a function of $\|\widetilde{\mathbf{A}}^\top \widetilde{\boldsymbol{\varphi}}_Y\|_\infty$ and $\|\bar{\mathbf{A}}^\top \boldsymbol{\varphi}_Y\|_\infty$ for the GL-LS and GL-L1 approaches, respectively, with $\|\cdot\|_\infty$ denoting the l_∞-norm. The choice of λ may be suboptimal, however it is motivated by the objective of obtaining solutions with both approaches that are generalisable for all different scenarios tested. In the implementation of the GL-LS and GL-L1 approaches, the weights are initialized as $w_l = 1$, for $l = 1, \ldots, L$, and in the case of GL-L1, the reweighting process is performed twice, resulting in three repetitions of the ADMM scheme. For each scenario considered, a total number of 100 Monte Carlo realizations are simulated. With the exception of Sec. 2.5.4, all scenarios are simulated under anechoic conditions, and with the exception of Sec. 2.5.5, the simulated source signals correspond to speech-shaped noise, obtained by white Gaussian noise with a 16-th order linear prediction filter based on a male speech signal from [169].

In Sec. 2.5.1, the performance measures used to evaluate the proposed approaches are defined. In Sec. 2.5.2 and Sec. 2.5.3, we present simulations in which a single point source is present, whereas in Sec. 2.5.5 and Sec. 2.5.6, the performance is evaluated in scenarios with 2 and 3 sources, respectively. The parameters varied for each simulated setup are further explained in each corresponding subsection.

Table 2.1: Main simulation parameters for each subsection

Subsection	P	N	Grid resolution	SNR	λ
5.2	1	6	[1°, 5°, 10°, 15°, 20°, 30°, 40°]	[0 dB, 3 dB, 5 dB, 10 dB, 15 dB, 20 dB]	$0.1\|\widetilde{\mathbf{A}}^\top\widetilde{\boldsymbol{\varphi}}_Y\|_\infty$ (GL-LS) $0.1\|\bar{\mathbf{A}}^\top\boldsymbol{\varphi}_Y\|_\infty$ (GL-L1)
5.3	1	6	10°	10 dB	$0.1\|\widetilde{\mathbf{A}}^\top\widetilde{\boldsymbol{\varphi}}_Y\|_\infty$ (GL-LS) $0.1\|\bar{\mathbf{A}}^\top\boldsymbol{\varphi}_Y\|_\infty$ (GL-L1)
5.4	1	6	10°	10 dB	$0.1\|\widetilde{\mathbf{A}}^\top\widetilde{\boldsymbol{\varphi}}_Y\|_\infty$ (GL-LS) $0.1\|\bar{\mathbf{A}}^\top\boldsymbol{\varphi}_Y\|_\infty$ (GL-L1)
5.5	2	6	10°	10 dB	$0.005\|\widetilde{\mathbf{A}}^\top\widetilde{\boldsymbol{\varphi}}_Y\|_\infty$ (GL-LS) $0.005\|\bar{\mathbf{A}}^\top\boldsymbol{\varphi}_Y\|_\infty$ (GL-L1)
5.6	3	9	10°	10 dB	$0.005\|\widetilde{\mathbf{A}}^\top\widetilde{\boldsymbol{\varphi}}_Y\|_\infty$ (GL-LS) $0.005\|\bar{\mathbf{A}}^\top\boldsymbol{\varphi}_Y\|_\infty$ (GL-L1)

2.5.1 Performance measures

The DOA estimation is evaluated by computing the mean absolute error (MAE) between the estimated DOA and the true source DOA, for each point source p separately, and can be expressed as

$$\mathrm{MAE}_p = \frac{1}{N_r} \sum_{r=1}^{N_r} |\vartheta_p^r - \hat{\vartheta}_p^r|, \tag{2.41}$$

where N_r denotes the total number of Monte Carlo realizations and r is the realization index.

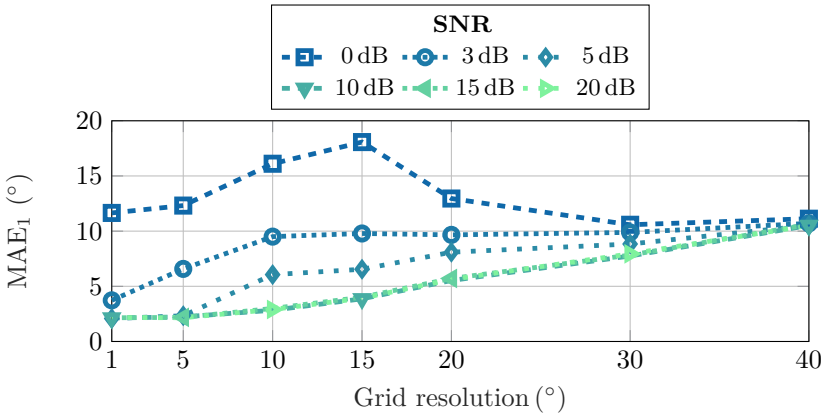
The point source PSD estimation is evaluated by computing the normalized mean squared error (NMSE), for each point source separately, as

$$\mathrm{NMSE}_p = \frac{1}{N_r} \sum_{r=1}^{N_r} \frac{\sum_{k=1}^{K} (\phi_{S_p}^r(k) - \hat{\phi}_{S_p}^r(k))^2}{\sum_{k=1}^{K} (\phi_{S_p}^r(k))^2}. \tag{2.42}$$

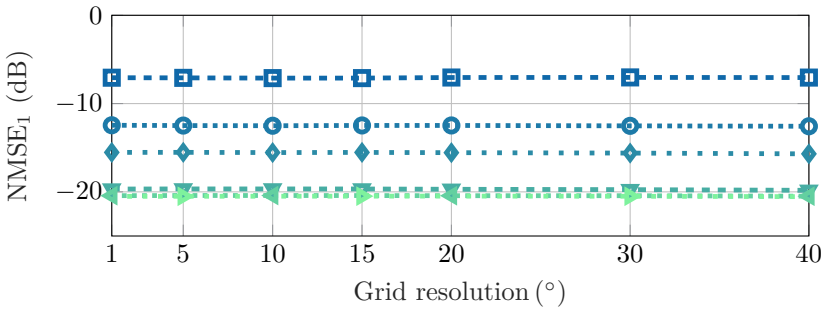
2.5.2 Different grid resolutions

Firstly, we consider a simulation setup for evaluating the performance of each method proposed as a function of the grid resolution selected for building the microphone response matrices, containing all candidate positions for a single point source. In addition, different SNR values are used to evaluate the methods' robustness to additive diffuse noise.

A single point source emitting stationary speech-shaped noise of variance $\sigma_{S_1}^2$ is simulated in anechoic conditions, with its DOA being randomly generated between 0° and 360° for each Monte Carlo realization, yielding the possibility of the source position being on or off-grid. A total of $N = 6$ observation frames are used, with the microphone orientations uniformly distributed over 360° (i.e., $0°, 60°, 120°, 180°, 240°$ and $300°$) for building the linear systems of equations, i.e., (2.16) and (2.12) for GL-LS and GL-L1, respectively. We assume the observations are made with an ideal cardioid microphone with flat frequency response, defined as $a_{\mathrm{cardioid}}(k, \theta) = 0.5 + 0.5\cos(\theta)$, $\forall k$, and that the microphone is static during each observation frame. The diffuse component is white Gaussian noise, with variance σ_D^2. The grid resolution is varied from 1° to 40°, and the SNR is varied from 0 dB to 20 dB. The regularization parameter λ is heuristically set to $0.1\|\widetilde{\mathbf{A}}^\top \widetilde{\boldsymbol{\varphi}}_Y\|_\infty$ and $0.1\|\bar{\mathbf{A}}^\top \boldsymbol{\varphi}_Y\|_\infty$ for GL-LS and GL-L1, respectively.



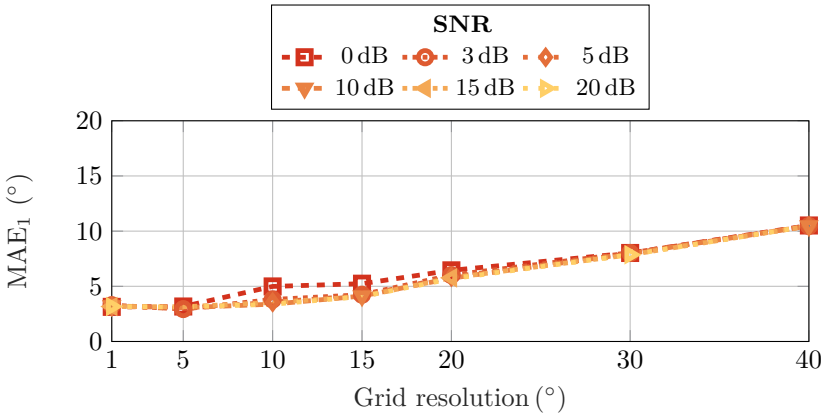
(a) DOA estimation error in terms of MAE



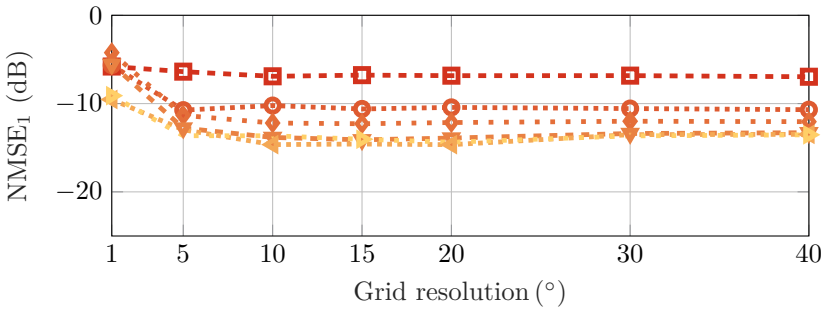
(b) PSD estimation error in terms of NMSE

Figure 2.1: Estimation errors obtained with GL-LS for different grid resolutions and SNRs

The estimation errors obtained when using GL-LS and GL-L1 for all combinations of grid resolution and SNR considered are presented in Fig. 2.1 and Fig. 2.2, respectively. When analyzing the DOA estimation in terms of MAE, we can observe that for both methods, the performance seems to converge to a certain minimum achievable error corresponding to a quarter of the grid resolution with an increase in SNR, which is a result of both methods picking the closest grid point to the point source's actual position. In that case, the MAE linearly increases as the grid resolution varies from 5° to 40°, whereas for the case of a one-degree resolution, the error indicates a possible limitation as a function of the angular variation of the cardioid directivity pattern in estimating the true source DOAs. For SNRs lower than 10 dB, the MAE obtained with the GL-LS



(a) DOA estimation error in terms of MAE



(b) PSD estimation error in terms of NMSE

Figure 2.2: Estimation errors obtained with GL-L1 for different grid resolutions and SNRs

approach varies more for different grid resolutions than the MAE obtained with GL-L1. This could indicate that while having a finer grid resolution can be beneficial in the DOA estimation, employing the narrowband system of equations where frequency-dependent information is preserved, as in the GL-L1 approach, can positively affect the robustness towards diffuse noise when trying to localize a point source that does not present a flat spectrum, such as the one simulated in this scenario.

When analyzing the PSD estimation in terms of NMSE, we observe that, for GL-LS, the NMSE seems to only depend on the SNR and not on the grid resolution, indicating that although the PSD re-estimation step via least squares depends on the previously estimated point source DOA, a mismatch between

the chosen angular grid for building the microphone response matrix and the source's true DOA does not impact the PSD estimation accuracy. This is also observed when GL-L1 is used with a grid resolution above 10°. When comparing both methods, we observe that GL-LS overall achieves a lower PSD estimation error than GL-L1 for different SNR values.

2.5.3 Different levels of microphone response mismatch

In this set of simulations, we aim to investigate the effect of a mismatch between the microphone response used for generating the input vectors $\widetilde{\boldsymbol{\varphi}}_Y$ and $\bar{\boldsymbol{\varphi}}_Y$ used by GL-LS and GL-L1, respectively, and the assumed microphone response when building the linear system of equations to be solved in each proposed approach. In practical scenarios, such a mismatch can occur when it is assumed that the microphone being used presents an ideally flat frequency response, whereas, in reality, it becomes more directional for higher frequencies instead. By fixing the use of an ideal cardioid microphone response when building the linear system of equations, a performance comparison between both approaches presented can be done for different cases of model mismatch caused by the use of the frequency-dependent microphone responses when generating the input vectors $\widetilde{\boldsymbol{\varphi}}_Y$ and $\bar{\boldsymbol{\varphi}}_Y$ used by GL-LS and GL-L1, respectively.

As an additional comparison, we also execute the proposed methods *without* including the additional nonnegativity constraint on the error term expressed in (2.17c) and (2.24c) for GL-LS and GL-L1, respectively, so that the possible improvement in robustness due to the constraint can be analyzed. In the case of GL-LS, this would correspond to solving the optimization problem presented in [121], and these versions of the proposed approaches are here referred to as GL-LS$_0$ and GL-L1$_0$.

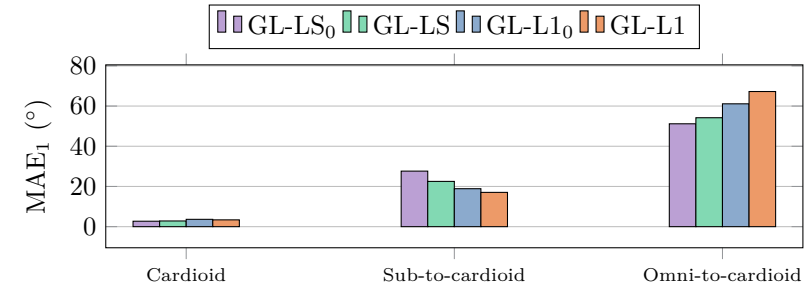
A single point source of speech-shaped noise is again simulated in anechoic conditions, with its DOA being randomly generated for each Monte Carlo realization. The SNR is here fixed at 10 dB and the grid resolution is fixed at 10°. For a normalized frequency value $f \in [0, 1]$, the frequency-dependent directivity patterns, denoted as *Sub-to-cardioid*[1] and *Omni-to-cardioid*, are defined as a linear combination of two directivity functions:

$$a(f, \theta) = (1 - f)a_L(\theta) + f a_H(\theta) \tag{2.43}$$

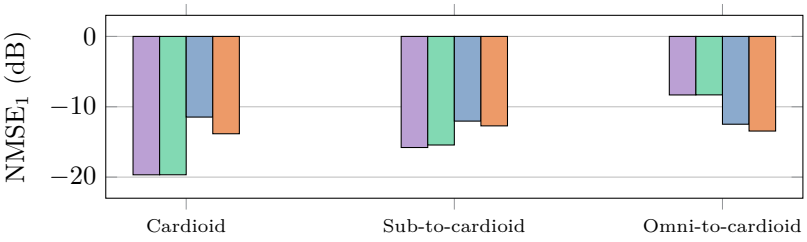
where:

$$\begin{aligned} a_H(\theta) &= 0.5 + 0.5\cos(\theta) \\ a_L(\theta) &= 0.75 + 0.25\cos(\theta) \end{aligned} \qquad \text{Sub-to-cardioid} \tag{2.44}$$

[1]The term '*Sub*' refers to the ideal subcardioid pattern, corresponding to the directivity function $a_L(\theta)$ in (2.44).



(a) DOA estimation error in terms of MAE



(b) PSD estimation error in terms of NMSE

Figure 2.3: Estimation errors obtained with GL-LS and GL-L1, as well as with their modified versions which exclude the nonnegativity constraint over the error term (GL-LS$_0$ and GL-L1$_0$, respectively), while using different microphone directivity patterns to generate the observed data

or:

$$\begin{aligned} a_H(\theta) &= 0.5 + 0.5\cos(\theta) \\ a_L(\theta) &= 1 \end{aligned} \qquad \text{Omni-to-cardioid} \tag{2.45}$$

The estimation errors obtained when using both versions of each proposed approach with different microphone responses are presented in Fig. 2.3. We can observe that for the case of generating data with an ideal cardioid, the additional constraint over the error term does not significantly affect the DOA estimation performance in terms of MAE for neither of the methods. However, it does result in slight improvement of the PSD estimation in terms of NMSE for the GL-L1 method. We can also observe that, when an actual mismatch between the microphone response used for building the response matrices and the microphone response used for generating the data is present, the DOA estimation is indeed improved for both approaches when considering the *Sub-to-cardioid* response, but not when considering the use of the *Omni-to-cardioid* response. This is

suspected to be due to low level of directivity in lower frequencies presented by the microphone response, such that the observed microphone PSD for different orientations does not provide sufficient directional information to appropriately localize the target source.

In terms of PSD estimation, we observe that the NMSE for GL-LS$_0$ and GL-LS do not significantly differ even with a more apparent difference in DOA estimation errors. This may be due to the fact that the GL-LS method is only affected by the additional nonnegativity constraint during the DOA estimation step, and the PSD is then re-estimated via least squares. In the latter step, the microphone response mismatch is still present and therefore not compensated for, possibly yielding similar error levels. A similar effect was observed in Sec. 2.5.2 for GL-LS, in which the mean mismatch between the true DOA and the chosen grid candidate, which depends on the grid resolution, did not impact the PSD estimation performance.

When comparing methods GL-L1 and GL-L1$_0$, the NMSE decreases with the inclusion of the additional nonnegativity constraint, even with an increase in MAE for the case of the *Omni-to-cardioid* microphone. This indicates that even if the additional constraint does not improve the DOA estimation, it can still positively affect the PSD estimation. We also observe that the GL-L1 approach seems to be more robust towards model mismatch than GL-LS in terms of PSD estimation, with or without the nonnegativity constraint over the error term, suggesting an advantage of employing the proposed narrowband system of equations instead of its wideband counterpart in this scenario.

2.5.4 Different levels of reverberation

While the simulations done in anechoic conditions in Sec. 2.5.2-2.5.3 were included to show important performance characteristics of the two proposed approaches, we here consider a more realistic scenario where a sound source is placed in reverberant environments.

A room of dimensions $6.3 \times 5.1 \times 2.5\,\text{m}$ is simulated, with a single, ideal cardioid microphone placed at coordinates $[3.7, 2.1, 1.5]\,\text{m}$. The point source is placed at the same height as the microphone and its DOA is set by randomly generating its coordinates within the room for each Monte Carlo realization, with the constraints of being at least $0.1\,\text{m}$ away from the room boundaries and exceeding the setup's critical distance from the microphone, denoted r_c, which varies with the reverberation time T_{60} considered [170]. An illustration of the simulated room with the constraints on the source position is shown in Fig. 2.4. For each of the $N = 6$ microphone orientations, the room impulse response is generated using the image source method implemented in [171] and convolved

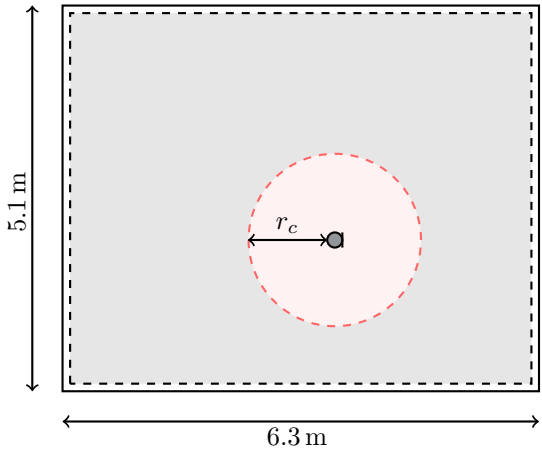
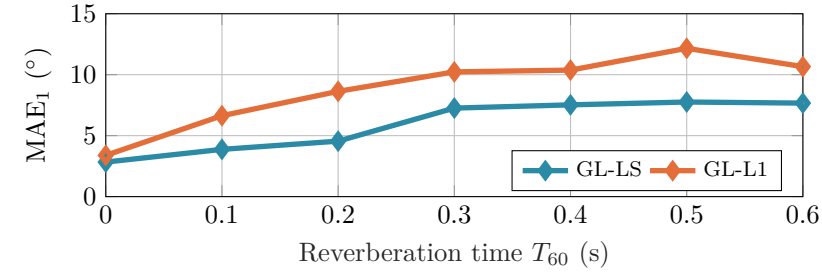


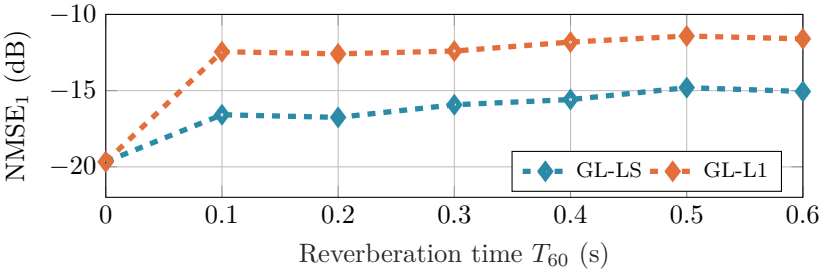
Figure 2.4: Illustration of simulated room where the source position is randomly generated within the area represented in gray (at least 10 cm from the wall and farther than the room's critical distance r_c)

with the original point source signal, composed of speech-shaped noise. The grid resolution for building the microphone response matrices is set to 10°, and the SNR is in this case defined as the ratio between the broadband power of the reverberant source signal and the diffuse noise component, with its value fixed at 10 dB. The reverberation time is varied from $T_{60} = 0$ s to $T_{60} = 0.6$ s, where $T_{60} = 0$ s corresponds to the anechoic case, and with reflections being simulated only on the two-dimensional plane in order to concord with the signal model in (2.1). Finally, the NMSE for evaluating the PSD estimation is computed with respect to a newly defined reference, corresponding to the PSD of the point source signal recorded with the cardioid microphone oriented towards the source's true DOA in the reverberant environment considered. This reference would correspond to the one expressed in (2.42) and used for evaluating the results obtained when considering anechoic conditions, as the reverberation time would correspond to zero.

The estimation errors obtained when using GL-LS and GL-L1 for all reverberation times considered are presented in Fig. 2.5. We can observe that the method GL-LS presents overall lower MAE and NMSE than the method GL-L1, indicating the benefit, in this scenario, of summing the signal PSD over frequency. Moreover, both approaches show the tendency of a performance degradation in DOA and PSD estimation with an increase in reverberation time, which is expected as reverberation is not explicitly accounted for in the utilized models (2.12) and (2.16).



(a) DOA estimation error in terms of MAE



(b) PSD estimation error in terms of NMSE

Figure 2.5: Estimation errors obtained with GL-LS and GL-L1 for different reverberation times

2.5.5 Influence of angular separation and power ratio between two sources

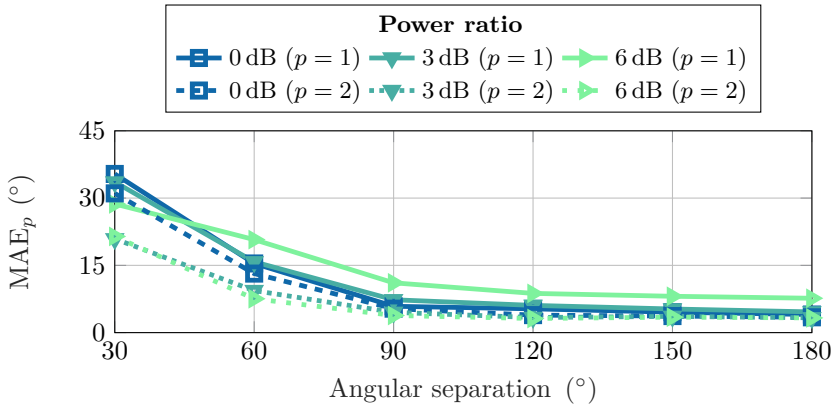
So far, we assumed that only a single point source was used when performing the simulations previously presented in Sec. 2.5.2, Sec. 2.5.3 and Sec. 2.5.4. Now, in order to allow a performance comparison between the proposed approaches regarding their capacity in separating different sources, we consider the case where two point sources with different spectral content are recorded by an ideal cardioid microphone. One of the point sources' DOA is randomly selected, whereas the DOA of the other source is then set according to a certain angular separation from the first source, varied from 30° to 180° with a 30°-step. The two sources, indexed as $p = 1$ and $p = 2$, emit colored noise based on a third-octave band filter centered on 1 kHz and 2 kHz, respectively. A power ratio between sources is defined as

$$\mathrm{PR} = \frac{\sigma_{S_2}^2}{\sigma_{S_1}^2}, \tag{2.46}$$

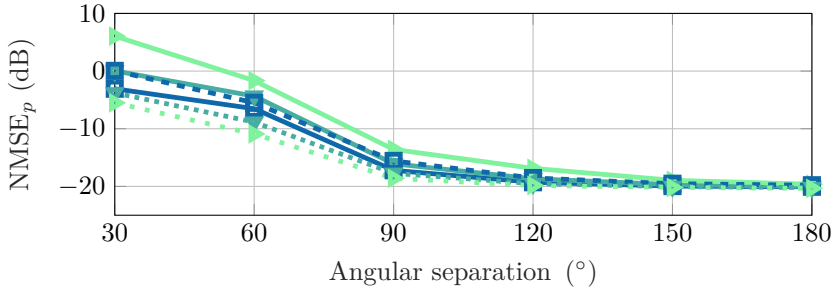
where $\sigma_{S_1}^2$ and $\sigma_{S_2}^2$ denote the broadband variance of sources $p = 1$ and $p = 2$, respectively. The power ratio is set to 0 dB, 3 dB and 6 dB according to (2.46). The grid resolution is fixed at 10° and the SNR, which is still defined with respect to the broadband variance of source $p = 1$, is fixed at 10 dB. The regularization parameter λ is now set to $0.005\|\widetilde{\mathbf{A}}^\top \widetilde{\boldsymbol{\varphi}}_Y\|_\infty$ and $0.005\|\bar{\mathbf{A}}^\top \boldsymbol{\varphi}_Y\|_\infty$ for the GL-LS and GL-L1 methods, respectively, as an attempt to decrease the influence of the group sparsity penalty, since in this scenario more than a single group is expected to be activated. The DOA and PSD estimation errors, computed for each source separately, are presented for the GL-LS and GL-L1 methods in Fig. 2.6 and Fig. 2.7, respectively. We can observe that, when using GL-LS, the MAE for both sources decreases as the angular separation between them increases, with the error being consistently lower for $p = 2$ when compared to $p = 1$ for PR = 3 dB and PR = 6 dB. We also observe that the difference in terms of MAE between the sources increases as the power ratio increases. These results are due to the second source having higher broadband power than the first, and therefore, being less corrupted by the diffuse noise in comparison. We also observe a similar behavior in the PSD estimation in terms of NMSE, with all errors converging to around −20 dB as the angular separation reaches 180°, and the difference in PSD estimation error between sources increasing with the power ratio.

When considering GL-L1 method, we can observe that, for angular separation values below 60°, its MAE is lower than the one obtained with GL-LS, indicating a benefit of using a frequency-dependent structure when building the proposed linear system of equations to identify closely located sources. However, as opposed to the behavior of GL-LS, the MAE for $p = 1$ increases with angular separation while the MAE for $p = 2$ remains reasonably constant for PR = 3 dB and PR = 6 dB. By further analysing the multiple realizations of each scenario considered, it was possible to observe that the estimated vector $\hat{\boldsymbol{\varphi}}_S$ often presented, especially for the case when PR > 0 dB, spurious peaks of frequency-averaged PSD neighboring the second source's estimated position ($\hat{\vartheta}_2$), indicating a spreading of the target source's power over multiple, neighboring candidate DOAs. If the frequency-averaged power of a candidate location neighboring source $p = 2$ is greater than the one related to source $p = 1$, then the algorithm will select the incorrect candidate and yield higher DOA estimation errors, which increase with the angular separation between sources and present opposing trends to those of the GL-LS approach. Since the regularization parameter has been heuristically chosen in this work's simulation and may be suboptimal, further tuning could potentially be carried out in order to improve this approach's DOA estimation performance in multi-source scenarios.

Despite an increase in MAE for $p = 1$ as a function of the angular separation, we can observe that, in terms of PSD estimation, the use of the GL-L1 method



(a) DOA estimation error in terms of MAE



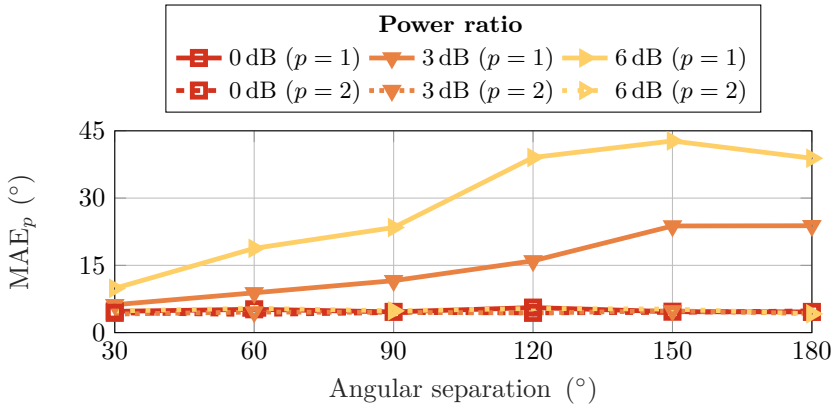
(b) PSD estimation error in terms of NMSE

Figure 2.6: Estimation errors obtained with GL-LS for different values of angular separation and power ratio between two sources

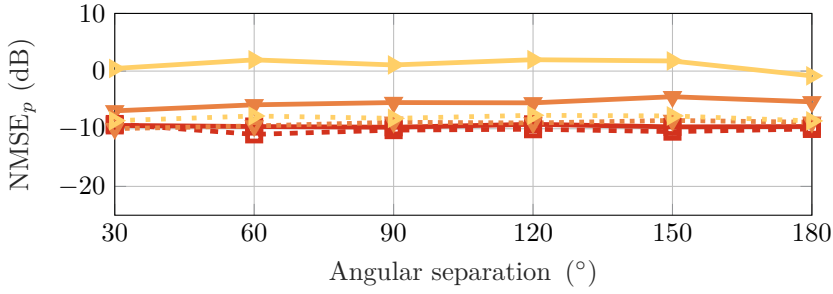
presents fairly constant NMSE values for both sources and all PR values considered. It is also observed that similarly to the use of GL-LS, the difference in terms of NMSE between two sources increases with the power ratio. Overall, the use of GL-LS showed to achieve lower PSD errors in most cases.

2.5.6 Influence of the microphone directivity pattern in the case of three sources

As a further investigation on the capacities of the proposed approaches to discriminate between different point sources in space, a new set of simulations is built for a case of three point sources. Upon testing the planned scenario, it



(a) DOA estimation error in terms of MAE



(b) PSD estimation error in terms of NMSE

Figure 2.7: Estimation errors obtained with GL-L1 for different values of angular separation and power ratio between two sources

was observed that using a cardioid microphone when simulating the recorded signals did not provide enough directional diversity to allow for three distinct peaks to be identified within the estimated vector $\hat{\boldsymbol{\varphi}}_S$. For this reason, we test both methods using microphones with higher-order directivity patterns based on the higher-order differential microphones studied in [172].

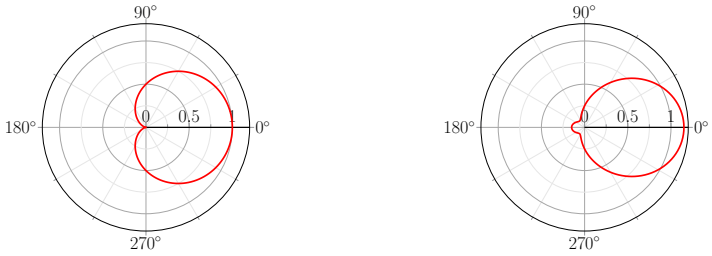
A general, second-order microphone directivity pattern, denoted $\Gamma(\theta)$, can be expressed as

$$\Gamma(\theta) = c_0 + c_1 \cos(\theta) + c_2 \cos^2(\theta) \tag{2.47}$$

where θ denotes the angle and c_0, c_1 and c_2 are real-valued scaling coefficients. By varying the values of c_0, c_1 and c_2, one can obtain different second-order directivity patterns. In this study, we consider simulating three different patterns,

Table 2.2: Coefficient values for microphone directivity patterns used

Directivity pattern	$[c_0, c_1, c_2]$
Cardioid	[0.5,0.5,0]
Cardioid-A	[0.25,0.5,0.4]
Cardioid-B	[0,0.5,0.5]
Hypercardioid-2	[-0.2,0.4,0.8]



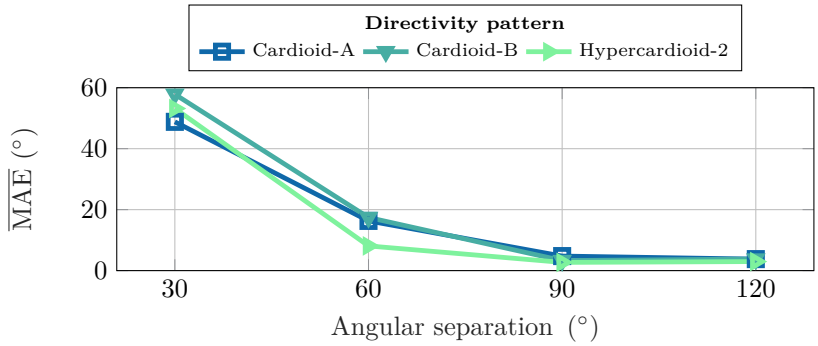
(a) Cardioid (b) Cardioid-A

(c) Cardioid-B (d) Hypercardioid-2

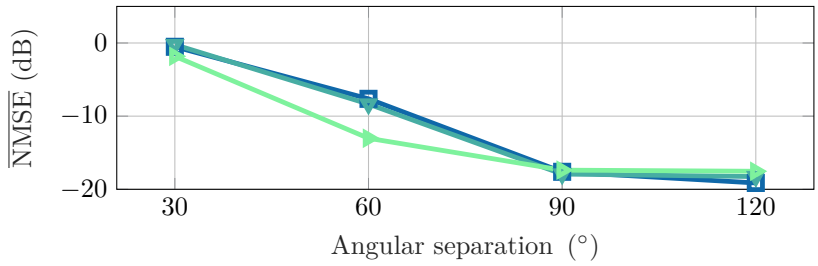
Figure 2.8: Microphone directivity patterns considered in this study represented in absolute values

here denoted *Cardioid-A*, *Cardioid-B* and *Hypercardioid-2*, based on the different combinations of coefficient values proposed in [172] and presented in Table 2.2. An illustration of each microphone directivity pattern in absolute values is presented in Fig. 2.8.

Three point sources of equal power emitting speech-shaped noise are simulated in anechoic conditions, with the DOA of the first source (ϑ_1) being randomly



(a) DOA estimation error in terms of $\overline{\text{MAE}}$

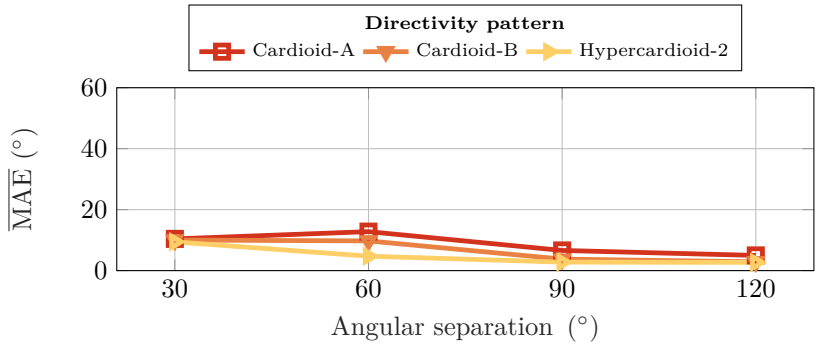


(b) PSD estimation error in terms of $\overline{\text{NMSE}}$

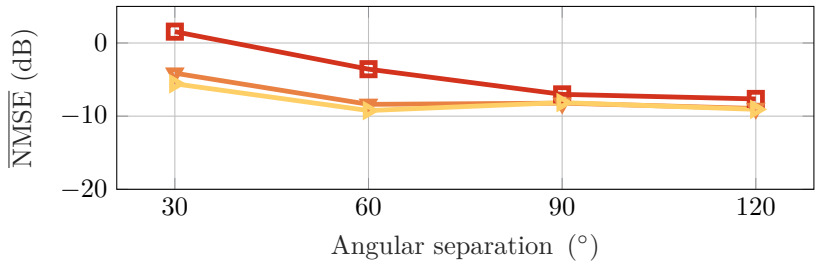
Figure 2.9: Estimation errors averaged over three sources and obtained with GL-LS for different values of angular separation, while using different microphone directivity patterns

selected between 0° and 360°, and the DOAs of the two remaining sources (ϑ_2 and ϑ_3) being set according to the angular separation considered as $\vartheta_2 = \vartheta_1 + \Delta\theta$ and $\vartheta_3 = \vartheta_1 + 2\Delta\theta$, with $\Delta\theta$ denoting such separation value. As opposed to the previous simulations presented, a total of $N = 9$ microphone orientations uniformly distributed over 360° (i.e., 0°, 40°, 80°, 120°, 160°, 200°, 240°, 280° and 320°) are used when building the linear system of equations used for each method, due to the need for more observation data in order to successfully differentiate the three different point sources. Both the DOA and PSD estimation errors are averaged over all three sources, since they are simulated with equal power, and are denoted as $\overline{\text{MAE}}$ and $\overline{\text{NMSE}}$, respectively.

The DOA and PSD estimation errors for the GL-LS and GL-L1 methods while using different second-order directivity patterns and angular separations between sources are presented in Fig. 2.9 and Fig. 2.10, respectively. When evaluating



(a) DOA estimation error in terms of $\overline{\text{MAE}}$



(b) PSD estimation error in terms of $\overline{\text{NMSE}}$

Figure 2.10: Estimation errors averaged over three sources and obtained with GL-L1 for different values of angular separation, while using different microphone directivity patterns

the DOA estimation, we observe again that the performance of GL-LS strongly depends on the angular separation between sources. However, the choice of microphone directivity pattern has only presented an impact for the case of an angular separation of 60° between sources, in which the *Hypercardioid-2* provided slightly better performance. For the GL-L1 approach, the same pattern yields lower or equal $\overline{\text{MAE}}$ values for all angular separations considered. We again observe that GL-L1 yields lower MAE than GL-LS for the case of sources with angular separation below 60°, indicating the benefit, in this case, of employing the narrowband signal model.

In terms of PSD estimation, it is observed again that, for both proposed approaches, the use of the *Hypercardioid-2* pattern overall yields $\overline{\text{NMSE}}$ values that are lower than or similar to those obtained with the *Cardioid-A* and *Cardioid-B* patterns. The GL-LS approach presents better PSD estimation

performance than GL-L1, despite its greater sensitivity to the angular separation between sources when performing the preceding DOA estimation step.

2.5.7 Discussion on the performance limitations of GL-LS and GL-L1

Based on the simulation results presented in this section, it can be observed that for both proposed approaches GL-LS and GL-L1, the DOA and PSD estimation performance can depend on multiple factors, which, when combined, can lead to the necessity of a thorough investigation over the conditions in which their application is to be considered.

In the single-source scenarios simulated in this work, the results indicate that although the intuitive choice of using a finer grid of candidate DOAs can aid at obtaining better DOA estimates, with the MAE being lowerbounded by approximately a quarter of the grid resolution, factors such as the noise level and different mismatches between the signal model assumed in the proposed approaches and the practical conditions in which the microphone signals are observed can strongly affect the overall performance of both DOA and PSD estimation. Regarding the model mismatches, it was observed that possible microphone calibration errors in its directivity and room reverberation lead to higher MAE and NMSE values.

In the multi-source scenarios, it was observed that the angular separation and difference in power between sources can significantly affect the DOA estimation performance of both proposed approaches and the PSD estimation performance of GL-LS.

Finally, although the extensive simulations presented in this work can already provide valuable information on the performance trends of the proposed approaches in numerous scenarios, more definitive evaluations can be obtained when considering the case of nonstationary signals and using experimental data. Many of the prevalent applications of DOA and PSD estimation involve speech signals and a combination of model mismatches resulting from multiple factors, which can be simultaneously present in practical setups. Therefore, an investigation of the proposed approaches' behavior under these conditions is required not only for gaining further clarity on the current applicability of the proposed approaches, but also to identify which main aspects should be considered in future work in order to enhance their performance.

2.6 Conclusion

In this paper, we proposed two approaches for performing DOA and PSD estimation of one or more point sources. The first approach, named GL-LS, is based on a broadband signal model with the PSDs summed over frequency for solving a group-sparse optimization problem with nonnegativity constraints over the desired output vector and the resulting error term, such that the sources' DOAs can be estimated from an overcomplete dictionary of angular candidate positions. Subsequently, a least squares step is performed for re-estimating the point sources' PSDs based on the estimated DOA information. The second approach, named GL-L1, is based on a narrowband signal model structure for solving an analogous optimization problem, which in this case is iteratively, group-wise reweighted for enhancing the solution's sparsity and jointly providing both DOA and PSD estimates.

Both approaches are implemented using ADMM, and simulations were performed for evaluating their performance under different conditions. Compared to the original method which GL-LS and GL-L1 were based on [121], it was observed that, in a scenario involving a microphone response model mismatch, having the additional nonnegativity constraint over the error term can improve the DOA estimation for the case of GL-LS and the PSD estimation for the case of GL-L1. Moreover, in terms of DOA estimation, the GL-L1 approach presented an advantage over GL-LS in scenarios with low SNR or where multiple sources are closely located to each other. Finally, it was shown that having the least squares PSD re-estimation step is beneficial in most scenarios, such that GL-LS outperformed GL-L1 in terms of PSD estimation errors.

Future work includes a further study of the influence of the choice of microphone orientations and directivity pattern when acquiring measurement data over the DOA and PSD estimation performance, expanding the proposed approaches to the multi-channel case and performing experimental tests.

Chapter 3

Multi-channel source localization

This chapter is based on the internal report:

Multi-source direction-of-arrival estimation using steered response power and group-sparse optimization

Elisa Tengan, Thomas Dietzen, Filip Elvander and Toon van Waterschoot

ESAT-STADIUS Tech. Rep. TR 23-163, KU Leuven, Belgium, submitted for publication, Dec. 2023.

Author contributions: All authors jointly developed the methodology and designed the computer simulations presented. ET implemented the algorithms and computer simulations, and all authors jointly interpreted the results obtained. ET drafted the manuscript and all authors read and reviewed the final manuscript.

3.1 Introduction

The field of sound source localization research has remained prevalent due to its wide-ranging applications, including domains as diverse as hearing aid technology [89, 130], multi-party videoconferencing [173], acoustic surveillance [10] and autonomous robots [101]. Over time, significant research efforts have been devoted to addressing the challenges of maintaining a robust source localization performance in realistic scenarios.

Many source localization methods rely on the use of microphone arrays, primarily due to the spatial diversity they offer and that can be effectively exploited [6]. Some well-established array processing methods for source localization include Capon's beamformer [15, 16], the MUSIC algorithm [20] and generalized cross-correlation (GCC) methods [13, 18, 19], while machine learning [132, 133] and compressed sensing techniques [41, 49, 50, 131] as well as methods based on optimal transport [174–176] have also emerged as relevant alternatives. Among the conventional GCC-based methods, the steered response power (SRP) has been thoroughly used as the base for many methods being further developed, mostly due to its robustness in adverse acoustic conditions such as noisy and reverberant environments [13, 18, 177]. Moreover, the use of the phase transform (PHAT) weighting on the GCC function yields the well-established SRP-PHAT framework, which has been demonstrated to improve performance on the localization of speech sources in reverberant environments [13, 18, 177–179].

Although the baseline SRP and SRP-PHAT methods are robust for single-source scenarios, numerous contributions have been made to address their limitations when confronted with the challenge of localizing multiple sources [44, 180–192]. Since the SRP and SRP-PHAT methods rely on finding the candidate direction that yields the maximum value over a spatial acoustic map [177], when multiple sources are present, differentiating true source-related peaks from spurious peaks caused by noise or acoustic reflections, as well as resolving closely located sources, becomes a difficult task [179, 188, 193].

Various strategies have been proposed to improve multi-source localization while employing an SRP-based framework. Some techniques propose the recursive removal of the influence of previously located sound sources in the observed SRP maps, thereby enabling the localization of the remaining sources [180–183]. Spatial clustering [184, 185] and grid refinement methods [186, 187] represent other relevant approaches in this context. Exploiting sparsity in the time-frequency domain [188–190] and in the acoustic scene [44, 191, 192] has also been investigated.

In this paper, we propose a method for estimating the direction of arrival (DOA) of multiple broadband sound sources in far-field scenarios based on a group-

sparse modeling framework. Group-sparse regularization, previously applied in the context of broadband source localization with co-prime arrays [156] and single directional microphones [121,124], is here employed to model a broadband SRP map. In particular, the cross-correlations used for computing the observed SRP map are modeled as a linear function of (unknown) source power spectral densities (PSDs), supported on an overcomplete search grid of candidate DOAs. Given the assumption of spatial sparsity, only a small number of candidate DOAs will have non-zero PSDs, yielding a group-sparse structure to the vector constructed by concatenating the DOA-specific PSDs from the entire search grid. We propose to estimate the full set of PSDs by solving a linear inverse problem augmented by a group-sparsity penalty. The source DOAs are then identified by the peaks of the spatial power density, i.e., the direction-specific PSDs integrated over frequency.

The proposed method is based on our previous work [194], where it is assumed that the search grid of candidate DOAs used to compose the PSD vector to be estimated coincides with the search grid of the observed SRP map that is being modeled. In this paper, we expand the methodology by reformulating the optimization problem to accommodate scenarios where the search grid used for computing the SRP map can have a lower resolution than the one used in the estimation of the source PSD vector. Within this context, the uniform spatial sampling requirements of the broadband SRP function, for a given microphone array configuration, are also derived in this work. An analytical expression on the sampling requirements of the GCCs used in the computation of the SRP map has been previously presented in [195], in terms of time difference of arrival (TDOA), and therein used to determine the pair-specific maximum signal bandwidth for which the sampling of the GCC function can be performed without aliasing, given a chosen grid resolution. Moreover, different approaches aimed at efficient spatial sampling of the SRP have also been proposed, based on the spatial distribution of TDOAs [196–198]. Alternatively, our focus lies in establishing, through a different perspective, a reference for the required uniform spatial grid resolution, and consequently, the number of required broadband SRP samples, to ensure the highest feasible spatial resolution in the observed SRP map. Furthermore, in the estimation of the PSD vector, a group square-root Lasso [199] formulation is used, as opposed to the more conventional group Lasso formulation [160] employed in our previous work [194]. As shown in [200], the group square-root Lasso is equivalent to a group-sparse covariance fitting problem, and its formulation has shown to facilitate the required regularization tuning based on group-sparse covariance fitting criteria [199,200]. This adaptation is particularly valuable in practical applications where the acoustic noise conditions are not known a priori. An efficient implementation of the proposed framework using the alternating direction method of multipliers (ADMM) [122] is also presented.

In order to evaluate the proposed method's DOA estimation performance, a series of simulations subject to different acoustic conditions are performed. We first consider a two-speaker scenario and subsequently vary the levels of noise and reverberation, as well as the angular separation and broadband power level difference between sources. We then consider scenarios with up to five sound sources, involving speech and instrumental music. Additionally, the effect of the difference between the search grid resolutions of the observed SRP map and the PSD vector is analyzed. These simulations greatly extend the aspects considered for evaluating the proposed method when compared with those presented in our previous work [194]. Moreover, the proposed method is compared with the conventional SRP-PHAT method, as well as with the group-sparse iterative covariance-based estimation (GSPICE) method, here used for estimating the PSDs directly from the microphone array cross-correlations, without computing an SRP map. Results show that the proposed method outperforms or performs comparably to the conventional SRP-PHAT method in most multi-source scenarios tested. Notably, even when the grid resolution of the modeled SRP map is lower than that of the estimated PSD vector, the proposed method shows it can maintain its performance levels. Furthermore, it is shown that the proposed method behaves overall similarly to GSPICE, at a lower computational cost, and that it particularly outperforms both SRP-PHAT and GSPICE in the cases of closely located sources.

This paper is organized as follows. In Sec. 3.2, the signal model is defined. In Sec. 3.3, the broadband SRP formulation, as well as its spatial sampling requirements, are detailed. In Sec. 3.4, the proposed method is presented. In Sec. 3.5, the simulation setups and their implementation are detailed, with results discussed in Sec. 3.6. Finally, Sec. 3.7 concludes the paper.

Notation: vectors and matrices are denoted by lower- and upper-case boldface letters, respectively. For matrices $\mathbf{A}$ (and vectors $\mathbf{a}$ where applicable), $\mathbf{A}^\top$, $\mathbf{A}^\mathrm{H}$, $\Re\{\mathbf{A}\}$, and $[\mathbf{A}]_{m,n}$ respectively denote the transpose, the Hermitian transpose, the real part, and the element of $\mathbf{A}$ at index (m, n). For vectors $\mathbf{a}$, $\|\mathbf{a}\|_p$ denotes the l_p-norm, the operation $\mathrm{diag}(\mathbf{a})$ creates a diagonal matrix from the elements of $\mathbf{a}$, and we use $\hat{\mathbf{a}}$ to denote an estimate of $\mathbf{a}$. For scalars a, the expectation operator is denoted by $\mathbb{E}\{a\}$, the complex conjugate by a^* and an estimate by $\hat{a}$. The operation $\mathrm{mod}(a, b)$ denotes a modulo b, and $\lceil a \rceil$ denotes the ceiling of a. Euler's number is denoted by e and the imaginary unit by j.

3.2 Signal model

We here consider scenarios where sound sources are located in the far field and in the same horizontal plane as a planar microphone array with M elements, indexed by $m = 1, \ldots, M$. The time difference of arrival (TDOA) of a propagating sound wave arriving from an azimuth angle γ that is observed at a microphone pair (m, m') is given by [13]

$$\Delta t_{m,m'}(\gamma) = (\mathbf{p}_m - \mathbf{p}_{m'})^\top \mathbf{r}_\gamma / c. \tag{3.1}$$

Here, $\mathbf{p}_m \in \mathbb{R}^3$ and $\mathbf{p}_{m'} \in \mathbb{R}^3$ are the Cartesian coordinate vectors of microphones m and m', respectively, i.e., $\mathbf{p}_m = [x_m, y_m, z_m]^\top$ and $\mathbf{p}_{m'} = [x_{m'}, y_{m'}, z_{m'}]^\top$. Furthermore, $\mathbf{r}_\gamma$ denotes the incident direction vector of unit length for azimuth angle γ, and c is the speed of sound. The azimuth angle of a sound source will also be referred to as its DOA.

In the discrete Fourier transform (DFT) domain, the m-th microphone signal is denoted by $y_m(\omega_k)$, with ω_k denoting the discrete angular frequency. Assuming that the microphone signals are sampled in the time domain with a sampling rate f_s, and that the signal $y_m(\omega_k)$ is computed from a fast Fourier transform (FFT) of length $2K$, the discrete frequencies correspond to $\omega_k = k\pi f_s / K$, with $k \in \{-K+1, \ldots, K\}$. Moreover, while considering a total of J sound sources located at DOAs $\vartheta_1, \ldots, \vartheta_J$, the signal $y_m(\omega_k)$ is modeled as

$$y_m(\omega_k) = \mathbf{w}_m^{\mathrm{H}}(\omega_k)\mathbf{s}(\omega_k) + n_m(\omega_k), \tag{3.2}$$

where $n_m(\omega_k)$ denotes the undesired component, which can be due to noise and reverberation. The vector $\mathbf{s}(\omega_k) \in \mathbb{C}^J$ corresponds to the source signal vector at the array's first microphone, here used as reference:

$$\mathbf{s}(\omega_k) = [s_1(\omega_k),\ \ldots\ ,\ s_J(\omega_k)]^\top, \tag{3.3}$$

with $s_i(\omega_k)$ denoting the signal emitted by the i-th source. The vector $\mathbf{w}_m(\omega_k) \in \mathbb{C}^J$ is thus the transfer function relative to the array's first microphone:

$$\mathbf{w}_m(\omega_k) = \left[e^{j\omega_k \Delta t_{m,1}(\vartheta_1)},\ \ldots\ ,\ e^{j\omega_k \Delta t_{m,1}(\vartheta_J)}\right]^\top, \tag{3.4}$$

such that, based on (3.3), $\mathbf{w}_1(\omega_k)$ corresponds to $\mathbf{w}_1(\omega_k) = [1, \ldots, 1]^\top$.

The generalized cross-correlation between recorded signals from a microphone pair (m, m') in the DFT domain is defined as

$$\psi_{m,m'}(\omega_k) = \mathbb{E}\{y_m(\omega_k) y_{m'}^*(\omega_k)\}. \tag{3.5}$$

If the source signals are assumed to be mutually uncorrelated, then the cross-correlation between $s_i(\omega_k)$ and $s_{i'}(\omega_k)$, for $i, i' \in \{1, \ldots, J\}$, is given by

$$\mathbb{E}\{s_i(\omega_k)s_{i'}^*(\omega_k)\} = \begin{cases} \phi_{S_i}(\omega_k) & \text{if } i = i' \\ 0 & \text{if } i \neq i' \end{cases} \tag{3.6}$$

where $\phi_{S_i}(\omega_k)$ is the power spectral density (PSD) of the i-th source signal. Furthermore, from (3.1), one can note that $\Delta t_{m,m'}(\gamma) = \Delta t_{m,1}(\gamma) - \Delta t_{m',1}(\gamma)$. Thus, the cross-correlation $\psi_{m,m'}(\omega_k)$ can be expressed as

$$\psi_{m,m'}(\omega_k) = \sum_{i=1}^{J} \phi_{S_i}(\omega_k) e^{-j\omega_k \Delta t_{m,m'}(\vartheta_i)} + \eta_{m,m'}^2(\omega_k), \tag{3.7}$$

where $\eta_{m,m'}^2(\omega_k)$ denotes the cross-correlation of the undesired component[1] for microphone pair (m, m').

3.3 Broadband steered response power (SRP) mapping

3.3.1 Broadband SRP formulation

For conciseness, a microphone pair (m, m') will be denoted by index p in the remainder, such that a given TDOA $\Delta t_{m,m'}(\gamma)$ will be denoted by $\Delta t_p(\gamma)$. For an M-element microphone array, the total number of different microphone pairs P is given by

$$P = M(M-1)/2. \tag{3.8}$$

Since $\psi_p(\omega_k)$ is Hermitian, we consider its single-sided form, with $k \in \{0, \ldots, K\}$. Assuming the DC component to be negligible ($\psi_p(\omega_0) = 0$), the broadband SRP for a given candidate DOA angle γ is defined as [201]

$$z(\gamma) = 2 \sum_{p=1}^{P} \sum_{k=1}^{K} \Re\{\psi_p(\omega_k) e^{j\omega_k \Delta t_p(\gamma)}\}. \tag{3.9}$$

[1] We note that there may be cross-correlation terms between the source signals and the undesired component, stemming from the latter's broad definition in (3.2). However, such cross-correlation terms do not impact the formulation of the optimization problem within the proposed method.

The exponential term in (3.9) corresponds to the phase shift applied to the pair-specific cross-correlation $\psi_p(\omega_k)$ in order to obtain a frequency-domain delay-and-sum beamformer steered towards candidate direction γ [18].

Considering a total of N look directions defining the SRP grid $\Gamma_N = \{\gamma_1, \ldots, \gamma_N\}$, N broadband SRP samples can be obtained by computing $z(\gamma_n)$ as in (3.9), where n denotes the look direction index. We define the broadband SRP map vector $\mathbf{z} \in \mathbb{R}^N$ as

$$\mathbf{z} = [z(\gamma_1),\ \ldots\ , z(\gamma_N)]^\top, \tag{3.10}$$

which can be expressed as a matrix-vector product as follows. We define, for a microphone pair p, the matrix $\mathbf{H}_p \in \mathbb{C}^{N\times K}$ and the vector $\boldsymbol{\psi}_p \in \mathbb{R}^K$, respectively, as

$$[\mathbf{H}_p]_{n,k} = e^{j\omega_k \Delta t_p(\gamma_n)}, \tag{3.11}$$

$$[\boldsymbol{\psi}_p]_k = \psi_p(\omega_k). \tag{3.12}$$

By stacking multiple matrices $\mathbf{H}_p$ and vectors $\boldsymbol{\psi}_p$, respectively, as

$$\mathbf{H} = [\mathbf{H}_1,\ \ldots\ , \mathbf{H}_P], \tag{3.13}$$

$$\boldsymbol{\psi} = \left[\boldsymbol{\psi}_1^\top,\ \ldots\ , \boldsymbol{\psi}_P^\top\right]^\top, \tag{3.14}$$

the broadband SRP map $\mathbf{z}$ can be expressed as

$$\mathbf{z} = 2\Re\{\mathbf{H}\boldsymbol{\psi}\}. \tag{3.15}$$

Conventionally, in a scenario with a single broadband source at DOA ϑ_1, performing the DOA estimation using an observed SRP map consists in finding the index n from the set of candidate DOAs that yields the maximum value of $z(\gamma_n)$:

$$\hat{n} = \arg\max_n z(\gamma_n), \tag{3.16}$$

such that the single source's estimated DOA is given by

$$\hat{\vartheta}_1 = \gamma_{\hat{n}}. \tag{3.17}$$

3.3.2 Uniform spatial sampling requirements

In this section, we describe the uniform spatial sampling requirements of the SRP function defined in (3.9). These requirements are presented with the

purpose of indicating the appropriate number of (spatial) SRP samples N to be computed, while ensuring the attainment of the maximum possible angular resolution over thee resulting SRP map $\mathbf{z}$. Note that the maximum resolution of the SRP beamformer relates to the minimum angular distance between sources such that they can be spatially distinguished, given the array configuration and signal bandwidth.

We start by analyzing the TDOA expression of a given microphone pair. As previously presented in (3.1), the TDOA depends on the incident direction vector $\mathbf{r}_\gamma$, which, in the case of far-field propagation and sound sources on the same horizontal plane as a planar microphone array, corresponds to [13]

$$\mathbf{r}_\gamma = -[\cos\gamma, \sin\gamma, 0]^\top. \tag{3.18}$$

The vector $(\mathbf{p}_m - \mathbf{p}_{m'})$ in (3.1) is given by

$$(\mathbf{p}_m - \mathbf{p}_{m'}) = \begin{bmatrix} x_m - x_{m'} \\ y_m - y_{m'} \\ 0 \end{bmatrix} = \begin{bmatrix} d_p \cos\alpha_p \\ d_p \sin\alpha_p \\ 0 \end{bmatrix} \tag{3.19}$$

where d_p denotes the Euclidean distance of the microphone pair as in $d_p = \|\mathbf{p}_m - \mathbf{p}_{m'}\|_2$, and α_p denotes the microphone pair's rotation on the horizontal plane. In Fig. 3.1, an example of a microphone pair on the horizontal plane and its corresponding distance d_p and rotation α_p are illustrated.

By decomposing the cross-correlation $\psi_p(\omega_k)$ into its magnitude and phase components $|\psi_p(\omega_k)|$ and $\angle\psi_p(\omega_k)$, respectively, the broadband SRP in (3.9) can be reformulated as

$$z(\gamma) = 2\sum_{p=1}^{P}\sum_{k=1}^{K} \Re\{|\psi_p(\omega_k)| e^{j\angle\psi_p(\omega_k)} e^{j\omega_k \Delta t_p(\gamma)}\} \tag{3.20a}$$

$$= 2\sum_{p=1}^{P}\sum_{k=1}^{K} |\psi_p(\omega_k)| \cos(\angle\psi_p(\omega_k) + \omega_k \Delta t_p(\gamma)), \tag{3.20b}$$

which corresponds to, essentially, a weighted sum of cosines with frequencies being dependent on $\Delta t_p(\gamma)$. Based on (3.1), (3.18) and (3.19), the SRP can be further reformulated as

$$z(\gamma) = 2\sum_{p=1}^{P}\sum_{k=1}^{K} |\psi_p(\omega_k)| \cos\left(\angle\psi_p(\omega_k) - \tfrac{\omega_k d_p}{c}\cos(\alpha_p - \gamma)\right). \tag{3.21}$$

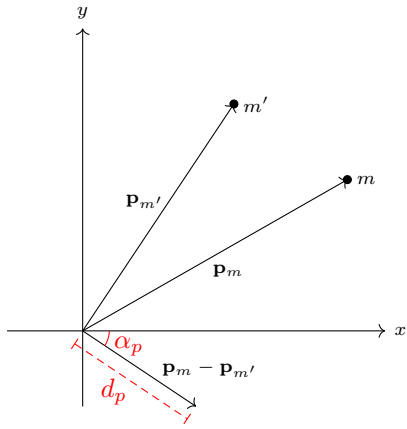


Figure 3.1: Example of a microphone pair on the horizontal plane with its corresponding inter-microphone Euclidean distance d_p and rotation α_p.

To obtain the approximate spatial frequency of $z(\gamma)$, we linearize the term $\cos(\alpha_p - \gamma)$ around an arbitrary look direction, denoted γ_0, by means of a first-order Taylor series approximation:

$$\begin{aligned}\cos(\alpha_p - \gamma) &\approx \cos(\alpha_p - \gamma_0) + \left(\frac{\mathrm{d}}{\mathrm{d}\gamma} \cos(\alpha_p - \gamma) \Big|_{\gamma=\gamma_0} \right) (\gamma - \gamma_0) \\ &= \cos(\alpha_p - \gamma_0) + \sin(\alpha_p - \gamma_0)(\gamma - \gamma_0),\end{aligned} \tag{3.22}$$

with $\frac{\mathrm{d}}{\mathrm{d}\gamma} \cos(\alpha_p - \gamma) \Big|_{\gamma=\gamma_0}$ denoting the derivative of $\cos(\alpha_p - \gamma)$ with respect to γ, evaluated in $\gamma = \gamma_0$. Hence, by inserting the above approximation in (3.21), we obtain

$$z(\gamma) = 2 \sum_{p=1}^{P} \sum_{k=1}^{K} |\psi_p(\omega_k)| \cos(\angle\psi_p(\omega_k) - \tfrac{\omega_k d_p}{c}(\cos(\alpha_p - \gamma_0) + \sin(\alpha_p - \gamma_0)(\gamma - \gamma_0))), \tag{3.23}$$

such that the spatial frequency of $z(\gamma)$ at look direction γ_0, for frequency ω_k and microphone pair p, is given by

$$\bar{\omega}_{k,p,\gamma_0} = \frac{\omega_k d_p}{c} \sin(\alpha_p - \gamma_0). \tag{3.24}$$

Based on the Nyquist-Shannon sampling theorem [202], and by assuming the microphone signals to be bandlimited by ω_b, i.e., $\psi_p(\omega_k) = 0$ for $\omega_k > \omega_b$, we

define, for microphone pair p, the corresponding critical spatial sampling period (in radians) as

$$\Delta\gamma_{p,\gamma_0} = \frac{\pi}{\max_{\omega_k \leq \omega_b} \bar{\omega}_{k,p,\gamma_0}} = \frac{\pi c}{\omega_b d_p \sin(\alpha_p - \gamma_0)}. \tag{3.25}$$

Note that the smallest value of $\Delta\gamma_{p,\gamma_0}$ for a given microphone pair p, denoted $\Delta\gamma_{p_{\min}}$, occurs at $\gamma_0 = \alpha_p - \pi/2$ and is equal to $\Delta\gamma_{p_{\min}} = (\pi c)/(\omega_b d_p)$. This observation is consistent with the one presented in [195]. Additionally, note that for a given microphone array configuration, the microphone pair with the largest inter-microphone distance, denoted $d_{\max}$, describes the smallest critical spatial sampling period over the entire array, corresponding to

$$\Delta\gamma = \frac{\pi c}{\omega_b d_{\max}}. \tag{3.26}$$

Therefore, if the SRP grid Γ_N used to compute the SRP map $\mathbf{z}$ is intended to be spatially uniform, one could consider $\Delta\gamma$ as an upper bound for the angular distance between grid points, which guarantees that $z(\gamma)$ is sufficiently sampled over γ for all microphone pairs and look directions. In that case, the minimum number of required SRP samples $N_{\min}$, uniformly distributed over 2π, is given by

$$N_{\min} = \left\lceil \frac{2\pi}{\Delta\gamma} \right\rceil = \left\lceil \frac{2\omega_b d_{\max}}{c} \right\rceil. \tag{3.27}$$

The knowledge of $N_{\min}$ is particularly relevant for compact microphone arrays. It has the potential to indicate whether the number of SRP samples to be considered, and therefore, the computational complexity of SRP-based DOA estimation methods, can be reduced while maintaining the highest achievable spatial resolution of the SRP map, given the available array configuration.

3.4 Proposed method

The conventional SRP approach detailed in Sec. 3.3.1 is evidently straightforward when performing single-source localization. However, when considering the case of multiple sources, especially under adverse acoustic conditions, such approach can be subject to DOA estimation errors caused by the presence of spurious peaks in the observed SRP map, as well as to the inability to resolve closely located sources. Therefore, in the following section, we present a method that aims to further improve the DOA estimation of multiple sources while considering the aforementioned limitations of the conventional SRP approach.

3.4.1 Group-sparse fitting of a broadband SRP map

We define a candidate source PSD vector $\boldsymbol{\varphi}_{S,l} \in \mathbb{R}^K$ as

$$\boldsymbol{\varphi}_{S,l} = \left[\phi_S(\omega_1, \theta_l) \;\; \dots \;\; \phi_S(\omega_K, \theta_l)\right]^\top, \tag{3.28}$$

where $\phi_S(\omega_k, \theta_l)$ is the candidate source PSD with DOA θ_l for frequency ω_k. By considering a set of L candidate DOAs $\Theta_L = \{\theta_1, \dots, \theta_L\}$ uniformly distributed over 2π and stacking different $\boldsymbol{\varphi}_{S,l}$ vectors, for $l = 1, \dots, L$, and for L much greater than the number of sources ($L \gg J$), the overcomplete vector $\boldsymbol{\varphi}_S \in \mathbb{R}^{KL}$ of direction-dependent source PSDs is obtained as

$$\boldsymbol{\varphi}_S = \left[\boldsymbol{\varphi}_{S,1}^\top \;\; \dots \;\; \boldsymbol{\varphi}_{S,L}^\top\right]^\top. \tag{3.29}$$

In addition, we define a vector $\mathbf{a}_{p,l} \in \mathbb{C}^K$ composed of the frequency-domain phase shifts related to a specific microphone pair p and candidate direction θ_l as

$$\mathbf{a}_{p,l} = \left[e^{-j\omega_1 \Delta t_p(\theta_l)} \;\; \dots \;\; e^{-j\omega_K \Delta t_p(\theta_l)}\right]^\top, \tag{3.30}$$

such that we can obtain a matrix $\mathbf{A} \in \mathbb{C}^{KP \times KL}$ as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1} & \dots & \mathbf{A}_{1,L} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{P,1} & \dots & \mathbf{A}_{P,L} \end{bmatrix}, \tag{3.31}$$

where $\mathbf{A}_{p,l} = \operatorname{diag}(\mathbf{a}_{p,l})$. Based on (3.7), we model the cross-correlation vector $\boldsymbol{\psi}$ in (3.14) as

$$\boldsymbol{\psi} = \mathbf{A}\boldsymbol{\varphi}_S, \tag{3.32}$$

so that the broadband SRP vector in (3.15) can be approximated as

$$\mathbf{z} \approx 2\Re\{\mathbf{H}\mathbf{A}\boldsymbol{\varphi}_S\}. \tag{3.33}$$

Since the PSD vector $\boldsymbol{\varphi}_S$ is real-valued, such that $\Re\{\mathbf{H}\mathbf{A}\boldsymbol{\varphi}_S\} = \Re\{\mathbf{H}\mathbf{A}\}\boldsymbol{\varphi}_S$, and by defining $\mathbf{B} = 2\Re\{\mathbf{H}\mathbf{A}\}$, we propose to estimate the PSDs corresponding to the candidate DOAs, and thereby the source DOAs, by solving

$$\underset{\boldsymbol{\varphi}_S \geq 0}{\text{minimize}} \quad \frac{1}{2}\left\|\mathbf{z} - \mathbf{B}\boldsymbol{\varphi}_S\right\|_2 + \sum_{l=1}^{L} \lambda_l \left\|\boldsymbol{\varphi}_{S,l}\right\|_2 \tag{3.34}$$

where λ_l is a regularization parameter. The proposed optimization problem's formulation corresponds to a nonnegative group square-root Lasso problem [199]. The regularization term $\sum_{l=1}^{L} \lambda_l \left\| \boldsymbol{\varphi}_{S,l} \right\|_2$ promotes solutions with a group-sparse structure, i.e., only a few of the subvectors $\boldsymbol{\varphi}_{S,l}$ composing $\boldsymbol{\varphi}_S$ should be nonzero, which agrees with the assumption of having a limited number of point sources ($J \ll L$). By estimating $\hat{\boldsymbol{\varphi}}_S$, and consequently, $\hat{\boldsymbol{\varphi}}_{S,1}, \ldots, \hat{\boldsymbol{\varphi}}_{S,L}$, the source signals' DOAs are estimated by identifying the peaks on the spatial power density, obtained by integrating the PSD values over frequency for each subvector $\hat{\boldsymbol{\varphi}}_{S,l}$, as previously proposed in [194]. For a total of J sources, the estimated DOAs $\hat{\vartheta}_1, \ldots, \hat{\vartheta}_J$ are then chosen to be the candidate directions related to the J largest peaks. One may note that the grid of candidate DOAs Θ_L employed to compose $\boldsymbol{\varphi}_S$ is not required to be equal to the SRP grid Γ_N used to obtain the observed SRP map $\mathbf{z}$, as opposed to the approach previously proposed in [194].

The choice of formulating a group square-root Lasso problem as opposed to the more conventional group Lasso [160] relates to the possibility of tuning the regularization parameter λ_l independently from the noise level [199]. Under the conditions of homoscedastic noise, and of same-size groups in the definition of $\boldsymbol{\varphi}_S$, the reference value $\tilde{\lambda}_l$ according to which the optimal choice of λ_l is scaled is given by [199, 200]

$$\tilde{\lambda}_l = \sqrt{\frac{\left\| \left[\|\mathbf{b}_{(l-1)K+1}\|_2^2 \ \cdots \ \|\mathbf{b}_{lK}\|_2^2 \right]^\top \right\|_2}{N}}, \tag{3.35}$$

for $l = 1, \ldots, L$, with $\mathbf{b}_q$ denoting the q-th column of $\mathbf{B}$.

3.4.2 Group square-root Lasso implementation

In order to solve the problem in (3.34), we propose an implementation based on the alternating direction method of multipliers (ADMM) [122].

By using auxiliary variables, denoted $\mathbf{u}_1$ and $\mathbf{u}_2$, the problem in (3.34) can be reformulated as

$$\underset{\mathbf{u}_1, \mathbf{u}_2 \geq 0}{\text{minimize}} \quad \frac{1}{2} \left\| \mathbf{u}_1 \right\|_2 + \sum_{l=1}^{L} \lambda_l \left\| \mathbf{u}_{2,l} \right\|_2 \tag{3.36a}$$

$$\text{subject to} \quad \mathbf{u}_1 = \mathbf{z} - \mathbf{B}\boldsymbol{\varphi}_S \tag{3.36b}$$

$$\mathbf{u}_2 = \boldsymbol{\varphi}_S \tag{3.36c}$$

where $\mathbf{u}_{2,l}$ corresponds to the l-th group composing the vector $\mathbf{u}_2$, such that it can be represented as

$$\mathbf{u}_2 = \begin{bmatrix} \mathbf{u}_{2,1}^\top & \dots & \mathbf{u}_{2,L}^\top \end{bmatrix}^\top. \tag{3.37}$$

The augmented Lagrangian of the optimization problem in (3.36) is given by

$$\begin{aligned} \mathcal{L}_\mu(\boldsymbol{\varphi}_S, \mathbf{u}_1, \mathbf{u}_2, \mathbf{d}_1, \mathbf{d}_2) = \frac{1}{2}\left\|\mathbf{u}_1\right\|_2 + \sum_{l=1}^{L} \lambda_l \left\|\mathbf{u}_{2,l}\right\|_2 \\ + \frac{\mu}{2}\left\|\mathbf{z} - \mathbf{B}\boldsymbol{\varphi}_S - \mathbf{u}_1 - \mathbf{d}_1\right\|_2^2 \\ + \frac{\mu}{2}\left\|\boldsymbol{\varphi}_S - \mathbf{u}_2 - \mathbf{d}_2\right\|_2^2 \end{aligned} \tag{3.38}$$

where $\mathbf{d}_1$ and $\mathbf{d}_2$ are the (scaled) dual variables, and μ can be interpreted as a dual update step size [122].

For a given ADMM iteration, indexed as q, we define the following shorthands:

$$\boldsymbol{\zeta}_1(q) = \mathbf{z} - \mathbf{B}\boldsymbol{\varphi}_S(q+1) - \mathbf{d}_1(q), \tag{3.39}$$

$$\boldsymbol{\zeta}_2(q) = \boldsymbol{\varphi}_S(q+1) - \mathbf{d}_2(q), \tag{3.40}$$

such that the updates for each variable in (3.38) can be written as

$$\boldsymbol{\varphi}_S(q+1) = \underset{\boldsymbol{\varphi}_S}{\arg\min}\, \mathcal{L}_\mu(\boldsymbol{\varphi}_S, \mathbf{u}_1(q), \mathbf{u}_2(q), \mathbf{d}_1(q), \mathbf{d}_2(q)), \tag{3.41}$$

$$\mathbf{u}_1(q+1) = \underset{\mathbf{u}_1}{\arg\min}\, \frac{1}{2}\left\|\mathbf{u}_1\right\|_2 + \frac{\mu}{2}\left\|\boldsymbol{\zeta}_1(q) - \mathbf{u}_1\right\|_2^2, \tag{3.42}$$

$$\mathbf{u}_2(q+1) = \underset{\mathbf{u}_2 \geq 0}{\arg\min} \sum_{l=1}^{L} \lambda_l \left\|\mathbf{u}_{2,l}\right\|_2 + \frac{\mu}{2}\left\|\boldsymbol{\zeta}_2(q) - \mathbf{u}_2\right\|_2^2, \tag{3.43}$$

$$\mathbf{d}_1(q+1) = \mathbf{u}_1(q+1) - \boldsymbol{\zeta}_1(q), \tag{3.44}$$

$$\mathbf{d}_2(q+1) = \mathbf{u}_2(q+1) - \boldsymbol{\zeta}_2(q). \tag{3.45}$$

It can be readily verified that the updates for the primal variables are given by

$$\boldsymbol{\varphi}_S\big(q+1\big) = \left(\mathbf{B}^\top\mathbf{B} + \mathbf{I}_{KL}\right)^{-1}\left[\mathbf{B}^\top(\mathbf{z} - \mathbf{u}_1(q) - \mathbf{d}_1(q)) + \mathbf{u}_2(q) + \mathbf{d}_2(q)\right], \tag{3.46}$$

Algorithm 3: The proposed DOA estimation algorithm

initiate $\hat{\boldsymbol{\varphi}}_S = \mathbf{0}^{KL\times 1}$
initiate $j := 0$, $\mathbf{u}_2(0) = \hat{\boldsymbol{\varphi}}_S$,
  $\boldsymbol{\varphi}_S(0) = \mathbf{d}_2(0) = \mathbf{0}^{KL\times 1}$, $\mathbf{u}_1(0) = \mathbf{d}_1(0) = \mathbf{0}^{N\times 1}$
repeat {ADMM scheme}
  update $\boldsymbol{\varphi}_S(j+1)$ as in (3.46)
  update $\mathbf{u}_1\big(j+1\big)$ as in (3.47)
  for $l \leftarrow 1$ to L **do**
    update $\mathbf{u}_{2,l}$ as in (3.48)
  end for
  $\mathbf{d}_1(j+1) = \mathbf{u}_1(j+1) - \boldsymbol{\zeta}_1(j)$
  $\mathbf{d}_2(j+1) = \mathbf{u}_2(j+1) - \boldsymbol{\zeta}_2(j)$
  $j \leftarrow j+1$
until convergence
store $\hat{\boldsymbol{\varphi}}_S = \mathbf{u}_2(\text{end})$
estimate $\hat{\vartheta}_i$, for $i = 1, \ldots, J$, as the J largest peaks integrated over frequency in $\hat{\boldsymbol{\varphi}}_S$

$$\mathbf{u}_1\big(q+1\big) = \mathrm{T}\left(\boldsymbol{\zeta}_1(q), 1/\mu\right), \tag{3.47}$$

$$\mathbf{u}_{2,l}\big(q+1\big) = \mathrm{T}\left(\max\left(\boldsymbol{\zeta}_{2,l}(q), 0\right), \frac{\lambda_l}{\mu}\right), \tag{3.48}$$

where $\mathbf{I}_{KL}$ denotes a $KL \times KL$ identity matrix, and $\mathrm{T}(\mathbf{x}, \kappa)$ is defined as the soft-thresholding operator

$$\mathrm{T}(\mathbf{x}, \kappa) = \left[\max\left(1 - \frac{\kappa}{\|\mathbf{x}\|_2}, 0\right)\right]\mathbf{x}. \tag{3.49}$$

In (3.48), the update of $\mathbf{u}_2$ is computed group-wise, for $l = 1, \ldots, L$. After computing the updates of $\mathbf{d}_1$ and $\mathbf{d}_2$ as in (3.44) and (3.45), respectively, the whole iterative process is repeated until convergence [122]. The proposed method's entire framework is summarized in Algorithm 3. It may be noted that the per-iteration complexity is dominated by (3.46), which, if implemented by using a Cholesky factorization of $(\mathbf{B}^\top\mathbf{B} + \mathbf{I}_{KL})$, has a complexity of $\mathcal{O}(K^2L^2)$. Furthermore, if $N < KL$, then the per-iteration complexity is $\mathcal{O}(NKL)$.

3.5 Simulations

In order to evaluate the DOA estimation performance of the proposed method, a series of simulations is performed under different acoustic conditions. In

Table 3.1: Source signal types.

Source	Signal type	Section
S_1	speech (male speaker) [204]	3.6.1-3.6.7
S_2	speech (female speaker) [204]	3.6.1-3.6.7
S_3	speech (male speaker) [204]	3.6.7
S_4	speech (female speaker) [204]	3.6.7
S_5	instrumental music (guitar) [205]	3.6.7

Sec. 3.5.1, the parameters describing the different simulated acoustic scenarios are defined. In Sec. 3.5.2, the choice of the search grids of candidate DOAs Γ_N and Θ_L are detailed. In Sec. 3.5.3, the performance metric used to evaluate the DOA estimation, as well as the methods with which the proposed method is compared, are specified. The particular characteristics of each scenario and the results obtained are further detailed and discussed in Sec. 3.6.

3.5.1 Simulation setup

Different far-field propagation scenarios are simulated in a rectangular room with dimensions $6.3 \times 5.1 \times 2.9\,\mathrm{m}$. A uniform circular array with a radius of 5 cm and $M = 6$ microphones is considered to be placed on a horizontal plane and centered at Cartesian coordinates $[3.1, 2.7, 1.5]^\top\,\mathrm{m}$, with a total of $P = 15$ distinct microphone pairs, and the largest inter-microphone distance corresponding to the array's diameter ($d_{\max} = 0.1\,\mathrm{m}$). The image source method implemented in [203] is used for simulating the room acoustics, characterized by the reverberation time (T_{60}).

The simulated point sources are denoted by S_i, where i is the source index, and are placed in the same horizontal plane as the microphone array, at a radial distance of 1.5 m from the center of the array. Each source emits a randomly selected segment of a pre-determined signal type. The signal type of each source and the section in which it is present is detailed in Table 3.1. With the exception of Sec. 3.6.6, all sources are simulated with equal broadband power. Spatially white speech-shaped noise is generated by filtering white Gaussian noise with a 16-th order linear prediction filter obtained from a male speech signal [169]. With $\sigma^2_{S_i}$ denoting the broadband power of the i-th source's reverberant signal

at the array's first microphone, the signal-to-noise ratio (SNR) is defined as

$$\mathrm{SNR} = \frac{\sigma_{S_1}^2}{\sigma_n^2}, \tag{3.50}$$

where σ_n^2 denotes the broadband power of the spatially white component. With a sampling rate of $f_s = 16\,\mathrm{kHz}$, the simulated time-domain microphone signals have a duration of 3 s, and their short-time Fourier transform (STFT) is computed using a 512-point square-root Hann window and 50% overlap, resulting in a time-domain window with a length of 32 ms and a total of 187 frames. We assume that in all scenarios, the microphone signal's bandwidth is limited by $\omega_b = 2\pi f_b$, with $f_b = 8\,\mathrm{kHz}$. The cross-correlations in (3.5) are obtained by averaging the PHAT-weighted STFT cross-correlations over a total of N_f frames [3]. Considering the single-sided broadband SRP formulation in (3.9), a total of $K = 256$ frequency bins are used.

For each scenario, a total of $R = 100$ realizations are performed, and for each realization r, the first source's DOA $\left(\vartheta_1^{(r)}\right)$ is randomly generated over the range of 360° without constraining it to lie on the candidate DOA search grids Γ_N and Θ_L being used. The remaining sources' DOAs are then set as

$$\vartheta_i^{(r)} = \vartheta_{i-1}^{(r)} + \Delta\vartheta_{i,i-1}^{(r)}, \quad \text{for } i = 2, \ldots, J, \tag{3.51}$$

where $\Delta\vartheta_{i,i-1}^{(r)}$ denotes the angular separation between two neighboring sources for realization r. This separation is randomly selected within a specified range in each scenario considered, and an example illustrating the spatial configuration of two sources is presented in Fig. 3.2.

3.5.2 Search grids of candidate DOAs

The set of candidate DOAs Θ_L, used in the proposed method's optimization problem to define the candidate source PSD vector $\boldsymbol{\varphi}_S$, is generated as a search grid with a fixed resolution of 5°, which yields a total of $L = 72$ candidate DOAs, uniformly distributed over 360°. The SRP grid Γ_N, used to compute the observed SRP map $\mathbf{z}$, is defined as a subset of Θ_L ($\Gamma_N \subseteq \Theta_L$), such that the N SRP samples composing $\mathbf{z}$ are also uniformly distributed over 360°. Based on the sampling requirements determined in Sec. 3.3.2, we note that, due to the

[3] Since the weighting operation is not accounted for in the cross-correlation model presented in Sec. 3.2, the PSD vector estimated with the proposed method will be inherently weighted as well in this simulation setup. However, due to the DOA estimation being based solely on the location of the nonzero elements of the PSD vector, it is expected that the proposed method's overall performance should not be affected.

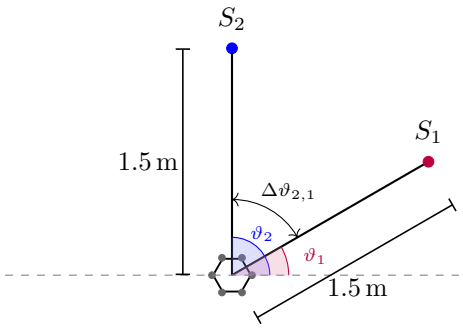


Figure 3.2: Illustration of a simulated scenario with two sources at a radial distance of 1.5 m from the center of a 6-element uniform circular array.

largest inter-microphone distance of the uniform circular array corresponding to its diameter ($d_{\max} = 0.1\,\text{m}$), the SRP map $\mathbf{z}$ should be composed, in theory, of at least $N_{\min} = 30$ uniformly distributed samples over the range of 360°, which corresponds to a grid resolution of approximately $\Delta\gamma \approx 12.2°$. Therefore, varying N allows for analyzing the different levels at which the possible spatial aliasing effects over the observed SRP map $\mathbf{z}$ can affect the DOA estimation performance.

3.5.3 Performance evaluation

The proposed method is implemented as in Algorithm 3, with the regularization parameter set to $\lambda_l = 0.01\tilde{\lambda}_l$, with $\tilde{\lambda}_l$ defined in (3.35), for $l = 1, \ldots, L$. The proposed method's DOA estimation performance is compared to that of the conventional SRP-PHAT method, and another group-sparse regression approach, namely the group-sparse iterative covariance-based estimation (GSPICE) method, proposed in [200] and briefly explained in the Appendix. The GSPICE method is here used to estimate the source PSD vector $\boldsymbol{\varphi}_S$, and consequently, the sources' DOAs, by directly fitting the microphone array cross-correlations vector $\boldsymbol{\psi}$ to the proposed source PSD model, according to (3.31). Therefore, one may note that the choice of the set of candidate DOAs Γ_N does not apply to GSPICE in this comparison. For all methods, the DOAs are estimated by finding the J largest local maxima on the corresponding maps obtained, and for the proposed method and GSPICE, the convergence criteria are defined with respect to the relative change in the estimated vectors between iterations, with a maximum number of iterations imposed.

Table 3.2: Main parameters for each set of simulations, divided by subsections

Section	N	N_f	$\Delta\vartheta_{i,i-1}^{(r)}$	SNR	T_{60}	J
3.6.1	12, 18, 24, 36, 72	64	$[30°, 60°]$	0 dB	0s	2
3.6.2	24, 72	1, 16, 32, 64, 128	$[30°, 60°]$	0 dB	0s	2
3.6.3	24, 72	64	$[5°, 15°]$, $[15°, 30°]$, $[30°, 60°]$, $[60°, 120°]$, $[120°, 180°]$	0 dB	0.2s	2
3.6.4	24, 72	64	$[30°, 60°]$	-18 dB to 18 dB	0.2s	2
3.6.5	24, 72	64	$[30°, 60°]$	0 dB	0s to 0.6s	2
3.6.6	24, 72	64	$[30°, 60°]$	0 dB	0.2s	2
3.6.7	24, 72	64	$[30°, 60°]$	0 dB	0s	1 to 5

The DOA estimation error for a specific source and realization is computed as

$$\epsilon_i^{(r)} = |\mathrm{mod}(\vartheta_i^{(r)} - \hat{\vartheta}_i^{(r)}, 360°)|, \quad \text{for } i = 1, \ldots, J, \tag{3.52}$$

and the accuracy score for evaluating the compared methods is defined as

$$\rho_i^{(r)} = \left(1 + e^{\sigma\left(\epsilon_i^{(r)} - \epsilon_{th}\right)/\epsilon_{th}}\right)^{-1}, \tag{3.53}$$

where ϵ_{th} corresponds to an absolute error threshold, in degrees. Here, ϵ_{th} is set to 5° in all cases and $\sigma = 6$. With the exception of Sec. 3.6.6, in which the DOA estimation accuracy is evaluated per source, the final score used for evaluation is computed by averaging the accuracy scores over all sources and realizations of the scenario being considered, as

$$\bar{\rho} = \frac{1}{JR} \sum_{i=1}^{J} \sum_{r=1}^{R} \rho_i^{(r)}, \tag{3.54}$$

and will be referred to as the average accuracy score in the remainder.

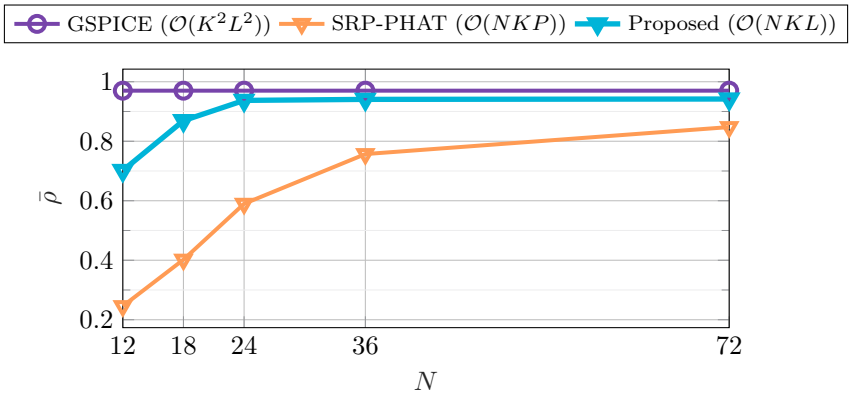
3.6 Results

In this section, we analyze the effect of each parameter described in Sec. 3.5.1 on the DOA estimation performance of the proposed method, in comparison with SRP-PHAT and GSPICE. In Table 3.2, a summary of the parameter values, used in each setup configuration being considered in the following subsections, is presented.

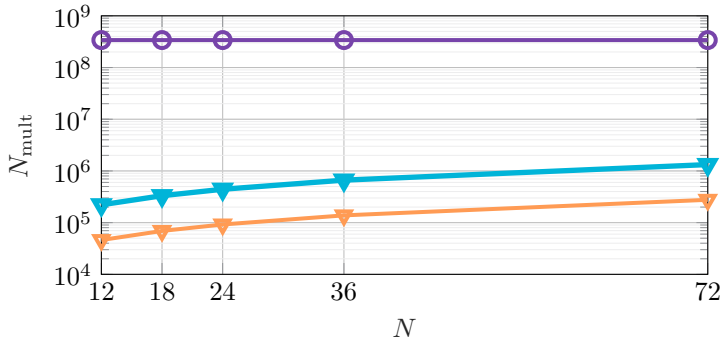
3.6.1 Number of SRP samples and computational complexity

Firstly, we aim to observe the effect of the choice of the candidate DOA search grid Γ_N used to compute the observed SRP map $\mathbf{z}$ in the DOA estimation performance of the proposed method and the conventional SRP-PHAT method. We also evaluate the computational complexity of the proposed method compared to SRP-PHAT and GSPICE. In this context, the computational complexity of SRP-PHAT is here determined by the cost of computing $\mathbf{z}$, which corresponds to $\mathcal{O}(NKP)$. For GSPICE, the per-iteration computational complexity corresponds to $\mathcal{O}(K^2L^2)$ [200], and for the proposed method, as previously mentioned in Sec. 3.4.2, it corresponds to $\mathcal{O}(NKL)$.

We consider a scenario with two sources (S_1 and S_2), with an angular separation of $\Delta\vartheta_{2,1}^{(r)} \in [30°, 60°]$. The SNR is set to 0 dB and the simulations are performed



(a) Average accuracy score $\bar{\rho}$ as a function of N.



(b) Number of per-iteration multiplications N_{mult} as a function of N.

Figure 3.3: Average accuracy score $\bar{\rho}$ and number of per-iteration multiplications N_{mult} as a function of N, for $K = 256$ and $L = 72$, with SNR $= 0\,\text{dB}$ and $\text{T}_{60} = 0\,\text{s}$.

under anechoic conditions ($\text{T}_{60} = 0\,\text{s}$). A total of $N_f = 64$ STFT cross-correlation frames are averaged in order to obtain the cross-correlation vector $\boldsymbol{\psi}$, and the number of points composing the SRP search grid Γ_N and used to compose the SRP map vector $\mathbf{z}$ is varied as $N = 12, 18, 24, 36, 72$, yielding the grid resolutions of $30°, 20°, 15°, 10°$ and $5°$, respectively.

In Fig. 3.3a, the average accuracy scores obtained with each method, and for the case of the proposed method and SRP-PHAT, with different values of N, are presented. It can be observed that, in this anechoic scenario, the proposed method overall outperforms SRP-PHAT. Moreover, the proposed method is

able to achieve a performance similar to the one obtained with GSPICE, with an increase in N. The proposed method's average accuracy score reaches its upper limit when $N = 24$, which corresponds to a grid resolution of 15° for the SRP grid Γ_N, and which is still below the minimum number of required SRP samples, $N_{\min}$. This indicates that while the knowledge of $N_{\min}$ can be beneficial for having a reference when designing the SRP grid Γ_N, such that aliasing should not occur in the SRP map, a robust performance can nonetheless still be achieved with the proposed method for lower grid resolutions of the SRP map $\mathbf{z}$ being employed. Given these observations, and for ease of interpretation, the simulation results in the following subsections are presented for $N = 24$ and $N = 72$.

In Fig. 3.3b, the computational complexity, in terms of the number of multiplications per iteration (N_{mult}), is presented for all three methods. We note that, in this set of simulations, the proposed method required on average 90 iterations to reach convergence, whereas GSPICE required on average 40 iterations. Nevertheless, the average runtime for the proposed method was lower than for GSPICE. When the complexity is evaluated alongside the average accuracy scores presented in Fig. 3.3a, we can observe that despite the proposed method being more computationally expensive than the conventional SRP-PHAT method, it still allows for an improvement in DOA estimation performance that is comparable to GSPICE, at a significantly lower computational cost per iteration.

3.6.2 Number of averaging STFT cross-correlation frames

In order to observe the effect of averaging over multiple STFT cross-correlation frames on the DOA estimation performance of the proposed method, we again consider the case of two sources (S_1 and S_2), with an angular separation of $\Delta\vartheta_{2,1}^{(r)} \in [30°, 60°]$. The SNR is set to 0 dB and the simulations are performed under anechoic conditions ($\mathrm{T}_{60} = 0\,\mathrm{s}$). The number of STFT cross-correlation frames being considered (N_f) is varied between 1 and 128. In Fig. 3.4, the average accuracy score obtained with the proposed method and SRP-PHAT for $N = 24$ and $N = 72$, as well as the average accuracy score obtained with GSPICE, is presented. It can be observed that in the two-source scenario considered, the performance of all methods improves as the number of STFT cross-correlation frames being averaged increases. We can also observe that the performance of SRP-PHAT is more dependent on the number of uniformly distributed SRP samples N than the proposed method, which exhibits a similar performance for both $N = 24$ and $N = 72$, as well as an increasing trend that is similar to that of GSPICE.

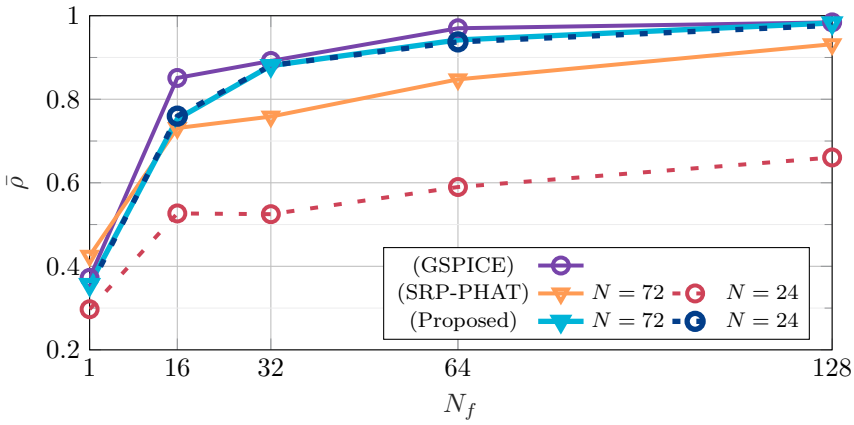


Figure 3.4: Average accuracy score obtained with GSPICE, SRP-PHAT and the proposed method for different numbers of averaging STFT cross-correlation frames N_f, with $N = 24$ and $N = 72$.

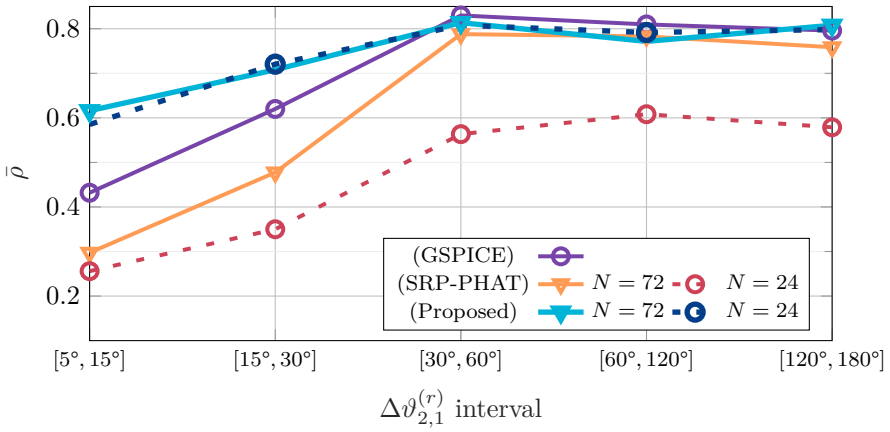


Figure 3.5: Average accuracy score obtained with GSPICE, SRP-PHAT and the proposed method for different angular separations between sources, with $N = 24$ and $N = 72$.

3.6.3 Angular separation between sources

In this set of simulations, we aim to investigate the influence of the angular separation between two sources (S_1 and S_2) of equal broadband power on

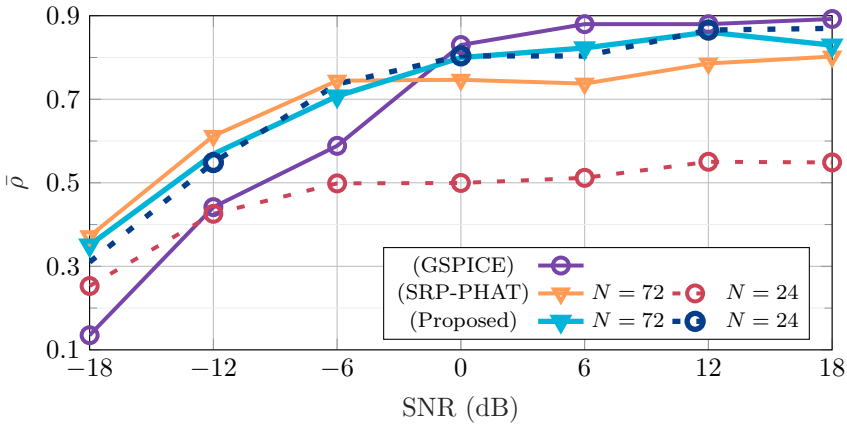


Figure 3.6: Average accuracy score obtained with GSPICE, SRP-PHAT and the proposed method for different SNRs, with $N = 24$ and $N = 72$.

the DOA estimation performance obtained with the proposed method, the conventional SRP-PHAT, and GSPICE. We vary the angular separation between the two sources such that, for each scenario considered, $\Delta\theta_{2,1}^{(r)}$ is within one of the following intervals: $[5°, 15°]$, $[15°, 30°]$, $[30°, 60°]$, $[60°, 120°]$, and $[120°, 180°]$. The SNR is set to 0 dB and $T_{60} = 0.2$ s. We consider the cases where $N = 24$ and $N = 72$ for the proposed method and SRP-PHAT, and the average accuracy scores obtained with each method are presented in Fig. 3.5. We again observe that the performance of SRP-PHAT is more dependent on the choice of N than the one obtained when employing the proposed method. We also observe that the DOA estimation performance overall improves as the angular separation between sources increases for all methods, with their best performance reached when the angular separation between sources is between 30° and 60° or greater. The proposed method performs particularly better than SRP-PHAT and GSPICE when $\Delta\theta_{2,1} \in [5°, 15°]$ and $\Delta\theta_{2,1} \in [15°, 30°]$, indicating its advantage in differentiating closely located sources.

3.6.4 Noise level

In this set of simulations, the effect of the spatially white noise level on the DOA estimation performance of the proposed method in comparison to the conventional SRP-PHAT and to GSPICE is analyzed by varying the SNR between −18 dB and 18 dB, for the case of two sources (S_1 and S_2) of equal broadband power. The reverberation time is set to $T_{60} = 0.2$ s and the angular

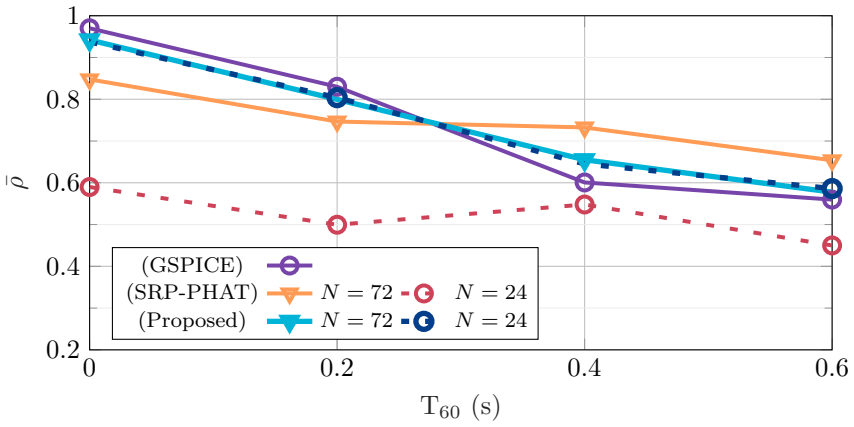


Figure 3.7: Average accuracy score obtained with GSPICE, SRP-PHAT and the proposed method for different reverberation times (T_{60}), with $N = 24$ and $N = 72$.

separation is $\Delta\vartheta_{2,1}^{(r)} \in [30°, 60°]$. The average accuracy scores obtained with each method for the different SNRs considered, as well as for $N = 24$ and $N = 72$ in the case of SRP-PHAT and the proposed method, are presented in Fig. 3.6. It can be observed that, as expected, the DOA estimation performance improves as the SNR increases for all methods being compared, with GSPICE exhibiting a greater sensitivity to the noise level than the proposed method and SRP-PHAT in cases of negative SNRs, while also exhibiting the best performance for positive SNRs. Moreover, we again observe that the proposed method performs consistently better than the conventional SRP-PHAT when $N = 24$. When $N = L = 72$, the proposed method still offers an advantage over SRP-PHAT for SNR values above 0 dB. Finally, we also observe that the proposed method's average accuracy score is similar for $N = 24$ and $N = 72$, indicating the ability to maintain performance while the search grid of candidate DOAs for composing the SRP map $\mathbf{z}$ can be coarser than the grid of candidate DOAs used for estimating the source PSD vector $\boldsymbol{\varphi}_S$.

3.6.5 Reverberation

The effect of reverberation over the DOA estimation performance of the proposed method in comparison to the conventional SRP-PHAT and to GSPICE is evaluated for the case of two sources (S_1 and S_2) of equal broadband power and with $\Delta\theta_{2,1} \in [30°, 60°]$. For SRP-PHAT and the proposed method, the

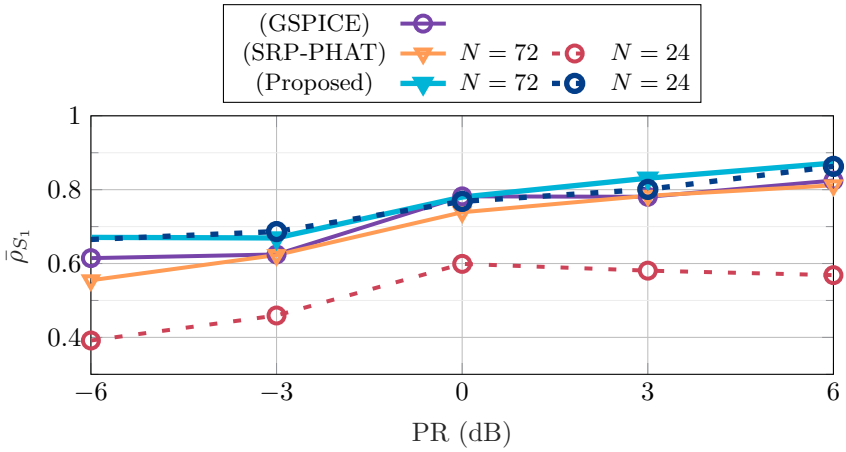
number of SRP samples is again set to $N = 24$ and $N = 72$. The SNR is set to 0 dB and the reverberation time is varied between 0 s and 0.6 s, with $T_{60} = 0$ s corresponding to the anechoic case. The average accuracy scores obtained with each method being compared are presented in Fig. 3.7, where it is observed that the proposed method, similarly to GSPICE, presents a DOA estimation performance that is seemingly more sensitive to an increase in T_{60} than the conventional SRP-PHAT, which could be due to the reverberation not being explicitly taken into account in the signal model used when formulating the proposed method. Nevertheless, for $N = 24$, the proposed method still performs better than SRP-PHAT for all reverberation times considered. When $N = 72$, SRP-PHAT presents an advantage in average accuracy score for $T_{60} \geq 0.4$ s. We again observe, as in the previous subsections, that the proposed method's average accuracy score obtained with $N = 24$ and $N = 72$ mostly coincide.

3.6.6 Broadband power ratio between two sources

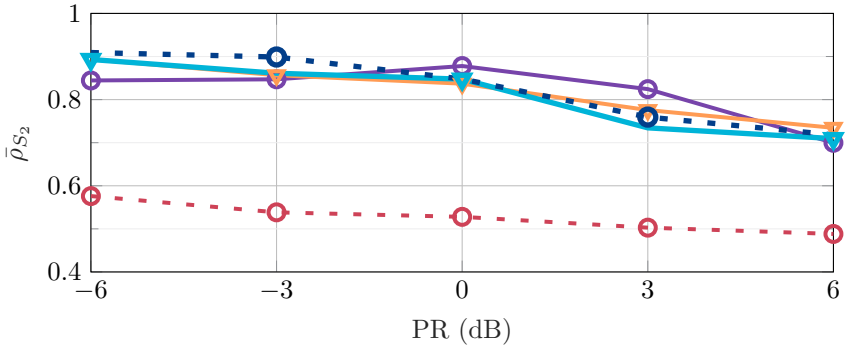
In order to investigate the effect of a possible power level difference between two sources, which commonly occurs in practical scenarios, on the DOA estimation performance of the methods being compared, we consider the case of two sources (S_1 and S_2) with $\Delta\vartheta_{2,1}^{(r)} \in [30°, 60°]$, and we define a broadband power ratio (PR) between the two sources as

$$\mathrm{PR} = \frac{\sigma_{S_1}^2}{\sigma_{S_2}^2}, \tag{3.55}$$

where $\sigma_{S_1}^2$ and $\sigma_{S_2}^2$ are defined as in Sec. 3.5.1. Since the speech sources S_1 and S_2 correspond to one male and one female speaker, respectively, and naturally present different spectral characteristics, the power ratio defined in (3.55) is not expected to be uniform over all frequencies. The power ratio PR is varied between −6 dB and 6 dB, the SNR is set to 0 dB and $T_{60} = 0.2$ s. Differently from the previous evaluations presented in Sec. 3.6.1-3.6.5, the average accuracy score is here computed per source and denoted by $\bar{\rho}_{S_1}$ and $\bar{\rho}_{S_2}$ for sources S_1 and S_2, respectively. The results per source obtained with $N = 24$ and $N = 72$ SRP samples are shown in Fig. 3.8. As the value of PR increases, we can observe that, overall, the average accuracy score for the first source increases, whereas it decreases for the second source. Moreover, we observe that, for all methods, the intersection between the average scores for each source does not occur exactly at PR = 0 dB, and that the performance range related to S_1 is slightly greater than the range related to S_2, for the PR values considered. This is most likely due to S_1 and S_2 being a male and female speaker, respectively, and consequently presenting different spectral characteristics. The proposed method presents consistently better localization performance for the male speaker (S_1) than the



(a) Average accuracy score with respect to S_1.



(b) Average accuracy score with respect to S_2.

Figure 3.8: Average accuracy score per source obtained with GSPICE, SRP-PHAT and the proposed method for different PRs, with $N = 24$ and $N = 72$.

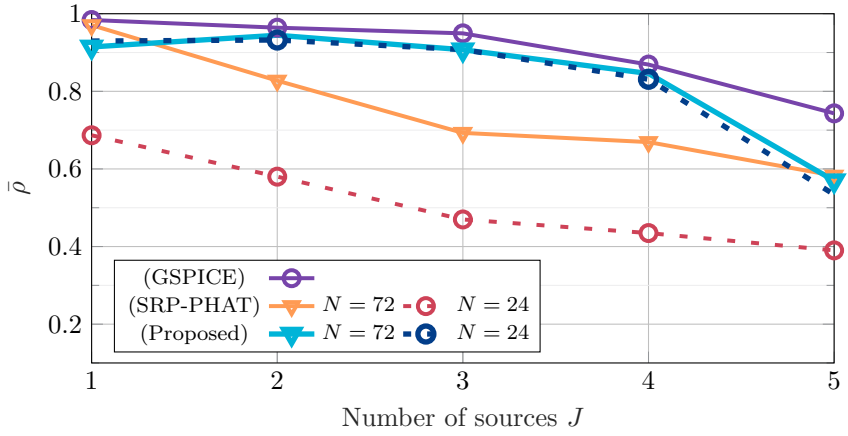


Figure 3.9: Average accuracy score obtained with GSPICE, SRP-PHAT and the proposed method for different numbers of sources, with $N = 24$ and $N = 72$.

conventional SRP-PHAT method for all PR values and for both $N = 24$ and $N = 72$. Regarding GSPICE, its performance for S_1 lies in between that of the proposed method and SRP-PHAT. When considering the female speaker (S_2), the proposed method's performance is better than the one obtained with SRP-PHAT for $N = 24$, and when $N = 72$, the proposed method performs similarly to SRP-PHAT only when PR $\leq 0\,\text{dB}$. In comparison to GSPICE, the proposed method presents an advantage in DOA estimation performance regarding S_2 for negative PR values.

3.6.7 Number of sources

Up until now, only two-source scenarios have been considered while evaluating the proposed method's DOA estimation performance in comparison with SRP-PHAT and GSPICE. In order to further investigate the proposed method's performance for more challenging scenarios, we now vary the number of sources present in the simulated acoustic scene from 1 to 5. The sources' DOAs are generated as explained in Sec. 3.5.1, with the angular separation between two subsequent sources $\Delta\vartheta^{(r)}_{i,i-1}$ set within the range of $[30°, 60°]$, while not being constrained to be equal between all neighboring sources. Each source is successively included in the simulated scenario in their corresponding order (S_1 to S_5). The SNR is set to $0\,\text{dB}$, and due to the increased number of sources in comparison to the previous sets of simulations, the reverberation time is set to $\text{T}_{60} = 0\,\text{s}$. The average accuracy scores obtained by the methods in

comparison and for $N = 24$ and $N = 72$ are shown in Fig. 3.9. We can observe that, under the anechoic conditions considered in this set of simulations, GSPICE presents overall the best DOA estimation performance. Although the proposed method does not present an advantage over SRP-PHAT for the single-source case when $N = 72$, it shows an improvement in DOA estimation in the multi-source cases involving up to 4 speech sources, with a performance trend somewhat similar to the one presented by GSPICE. When the fifth source is included in the simulation setup, the proposed method's performance is more significantly affected. Nevertheless, when $N = 24$, the proposed method outperforms SRP-PHAT in all single and multi-source scenarios tested.

3.7 Conclusion

In this paper, we proposed a method for estimating the DOAs of multiple broadband sources by solving a group-sparse optimization problem that models a broadband SRP map as a linear combination of PSDs, and where the grid resolution of the observed SRP map can be lower than the resolution of the DOA search grid used to model the candidate source PSDs. The estimation of the source DOAs is then accomplished by identifying peaks in the resulting spatial power density, i.e., the direction-specific PSDs integrated over frequency. An implementation using the alternating direction method of multipliers (ADMM) is presented, and simulation results showed that, when using SRP maps with lower grid resolutions with respect to the sampling requirements of the broadband SRP function, the proposed method outperformed the conventional SRP-PHAT method in the multi-source scenarios tested. It was also shown that the proposed method performed similarly to the group-sparse iterative covariance-based estimation (GSPICE) method, here applied to estimate the source DOAs by directly fitting the microphone array cross-correlations into the candidate source PSD model used, at a lower computational cost. Finally, it was observed that the proposed method outperformed both SRP-PHAT and GSPICE in the cases of closely located sources.

Part II

Speech enhancement in adverse acoustic conditions

Chapter 4

Experimental setup for measuring UAV noise

This chapter is based on the publication:

Drone noise directivity and psychoacoustic evaluation using a hemispherical microphone array

Mansour Alkmim, João Cardenuto, Elisa Tengan, Thomas Dietzen, Toon van Waterschoot, Jacques Cuenca, Laurent De Ryck and Wim Desmet

Published in *J. Acoust. Soc. Am.*, vol. 512, no. 5, p. 2735–2745, 2022.

Author contributions: MA, JC, ET and TD jointly designed the experimental setup and executed the experimental measurement campaign. MA drafted the manuscript, and all authors read and reviewed the final manuscript.

© 2022 The Authors. Reprinted, with permission, from:
M. Alkmim, J. Cardenuto, E. Tengan, T. Dietzen, T. van Waterschoot, J. Cuenca, L. De Ryck and W. Desmet, “Drone noise directivity and psychoacoustic evaluation using a hemispherical microphone array,” *J. Acoust. Soc. Am.,* vol. 512, no. 5, p. 2735–2745, 2022.
Changes include layout, representation, and editing.

4.1 Introduction

Vertical takeoff and landing (VTOL) unmanned aerial vehicles (UAVs), also known as drones, have been attracting great interest for many commercial and recreational applications such as point-to-point delivery, photography, and various monitoring activities. VTOL UAVs have lower noise emission levels compared to conventional aircraft [206]. However, an increase in the number of drones in operation for civil applications (i.e., low altitudes) will likely cause noise annoyance issues in dense urban centers [207]. Currently, in the European Union, VTOL UAV noise emissions are not formally regulated. Nevertheless, the general requirement of noise measurement of lightweight and small multirotor systems is under development [208], and the matter has been given high priority to ensure safe, secure, and sustainable operation [209, 210].

Due to their compact design, the noise emitted by drones has a distinct acoustic signature, induced mainly by the side-flow horizontal motion and rotor synchronization [207]. The dominant noise source is generated by the rotating blades at the blade passing frequency (BPF) — a frequency component relating to the number of blades and the rotation speed — and their harmonic components. The tonal noise is often considered the cause of annoyance. On the other hand, the broadband noise is generated by the electric motor and by turbulences in the flow [211]. Additionally, the emitted sound and radiation pattern can be affected by scattering effects on the drone body [212].

Drone noise emissions have been modeled both analytically [211] and numerically [212, 213]. These models are either simplified or require a high level of complexity and long computational time. Alternatively, controllable and reproducible measurements can provide useful insights and are essential to validate and improve the models. Besides that, the measured data can be directly used as input to the propagation and auralization models. An experimental aeroacoustic study of a drone for different rotor blades was performed by [214]. Measurements were performed with a single microphone 1.51 m away from the center of the drone at a 50° elevation with respect to the drone's horizontal plane. In [215], drone noise was synthesized using measurements of the radiation strength using five sensors at five elevations and using in-field recordings. The recordings were manipulated to simulate different maneuver conditions and changes in rotor speed and to include outdoor sound propagation effects. The synthesized audio was plausible but with some limitations such as flight dynamics and coupling effects between propellers, which were not taken into account. In [216], a Gaussian beam tracing method was proposed, which asymptotically approximates the wave equation to model the drone emission and propagation. The method allowed for generic source directivity patterns and a realistic urban environment.

A comprehensive literature review on drone noise emission and noise effects on humans can be found in [207]. As reported, most works on drone noise emission have been performed in laboratory conditions but with a limited aperture and number of sensors. Noise emission levels were investigated both as a function of takeoff mass and in terms of directivity. The latter is known to have a major influence and needs to be considered in the noise emission model. However, measurement procedures to capture

the full directivity of noise emissions of drones are still scarce. Hence, this chapter proposes an experimental protocol for measuring the noise emission and directivity of VTOL UAVs.

A hemispherical microphone array setup is developed for measuring the directivity pattern of the sound pressure generated by a drone operating at a simulated constant thrust. Different from previous works, the noise source can be captured simultaneously for different elevation and azimuth angles. The methodology is demonstrated using a quadrotor drone, and the analysis on the noise emission levels is performed for a fixed azimuth and varying elevation angles. The results obtained indicate that the sound pressure levels (SPLs), computed for all observed elevation angles, increase with the rotor speed. Moreover, when analyzing the SPLs in one-third octave bands, it is observed that although the drone noise emission levels are similar for different elevation angles, variations in their spectral characteristics can be detected.

This chapter is organized as follows. The measurement setup is described in Sec. 4.2, the results on drone noise emission levels are presented and discussed in Sec. 4.3, and Sec. 4.4 concludes the chapter.

4.2 Measurement setup

The measurements were performed in a semi-anechoic chamber with an acoustically hard ground as shown in Fig. 4.1. The room interior dimensions are $16 \times 12.5 \times 6\,\mathrm{m}$ (length $\times$ width $\times$ height) with fiberglass wedges covered by perforated aluminum panels. Additionally, a $2.5 \times 2.5\,\mathrm{m}^2$ area under the UAV was covered by a foam material with approximately 0.07 m thickness to mitigate ground reflections. During measurements, the temperature of the room ranged from 20°C to 23°C. The speed of sound is here considered as $c = 343\,\mathrm{m/s}$. The measured UAV corresponds to a MikroKopter MK EASY Quadro V3 (HiSystems GmbH, Moormorland, Germany) with a custom-made bottom frame. The UAV has four rotors in "+" configuration, each with a two-blade propeller, and weighs in total 2.5 kg. The propeller diameter is $d = 0.2748\,\mathrm{m}$, yielding a drone pitch of $\lambda = 2.18$ ($\lambda = D/d$, where D is the distance between two opposite rotors). The UAV was rigidly attached to a heavy support stand, blocking all six degrees of freedom. The spherical coordinate system, illustrated in Fig. 4.2, has its origin defined at the geometric center of the drone, which is at a height of 1.57 m from the floor. To capture the radiated pattern, a hemispherical array of 18 microphones was built with an average radius of $0.98 \pm 0.10\,\mathrm{m}$. The configuration of the microphones, as shown in Fig. 4.1, covers a hemispherical region with the center matching the drone geometric center. The microphones are 1/4" externally polarized integrated circuit piezoelectric microphones [GRAS (Holte, Denmark) 40-PH]. Microphone 1 was covered with a windscreen foam as it lies directly underneath the propellers. Fig. 4.3 shows a schematic of the microphone array configuration. For instance, microphones 1, 2, and 3 are at 0° azimuth and approximately at $-70°$, $-50°$, and $-20°$ elevation, respectively, with respect to the horizontal plane in the center

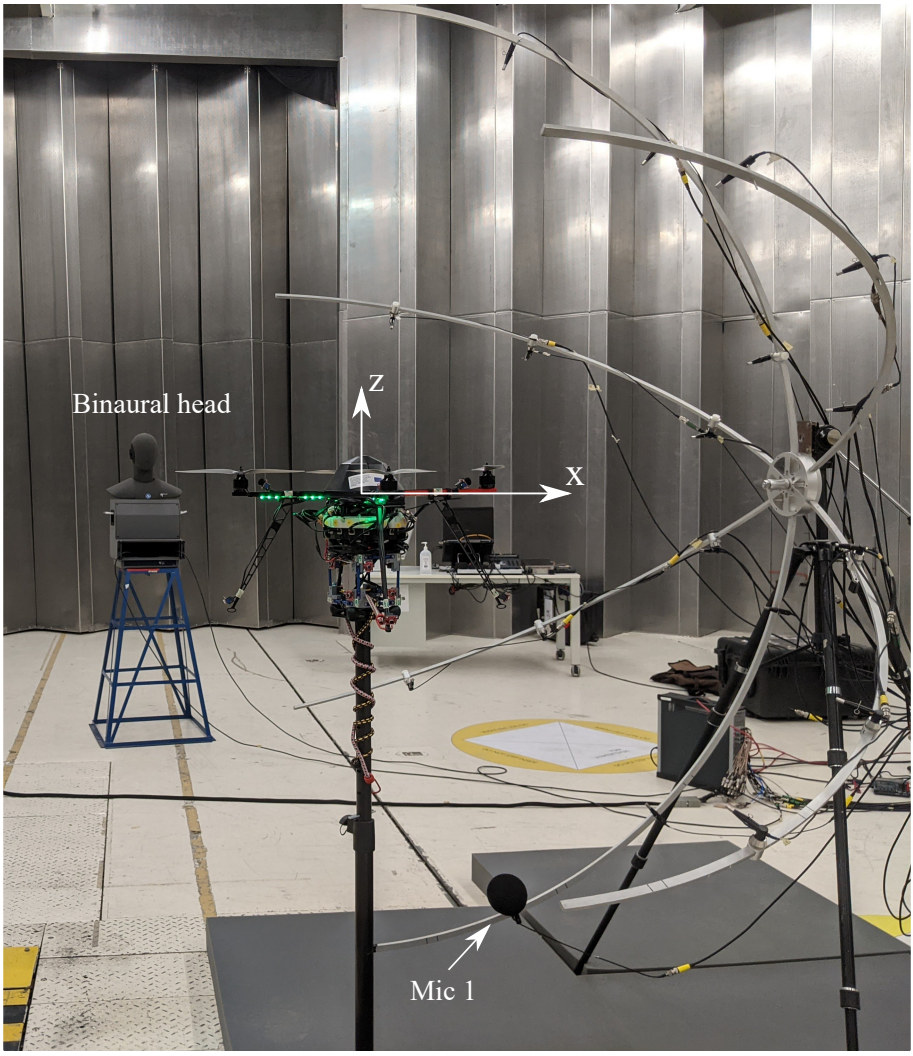


Figure 4.1: Experimental setup consisting of a hemispherical microphone array and a UAV mounted on a stand. The binaural head depicted in the background has not been used in the processing of signals in the results presented in Sec. 4.3.

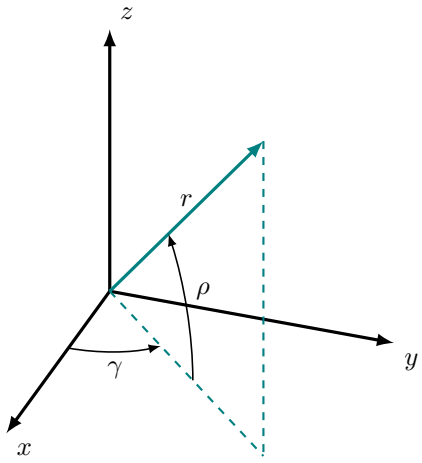


Figure 4.2: Spherical coordinate system, where γ denotes the azimuth, ρ the elevation, and r the radius.

of the drone. The numerical order of the microphone follows each arm of the array frame in a counterclockwise rotation starting from the bottom arm at 0° azimuth.

The microphone signals are sampled at 52.1 kHz. Additionally, a laser probe [KEYENCE (Osaka, Japan) FSV31P] with a detection time of 33 µs placed under one of the propeller blades to measure the rotational speed at run time. Measurements were performed in operational conditions by controlling the transmitter [Graupner (Kirchheim unter Teck, Germany) MC-20] throttle joystick. The transmitter allows small increments and stable control of the vertical thrust since the joystick is not spring-loaded. The radiation pattern can be determined for a fixed position with varying rotor speeds in rpm, which simulates a hovering condition for different payload capacities. Rotor interaction plays a role in sound generation, especially in forward motion and descent maneuvers. For instance, forward motion occurs when the rotational speed of the front rotors is different from the rear rotors and the UAV is slightly tilted forward. In the descent maneuver, the noise generation is more complex as it features fluctuations in sound pressure levels (SPLs) due to unsteady airflow conditions [217]. In this paper, the rotor interaction from maneuvers is not accounted for in the measurement as the drone is at a fixed orientation. Additionally, the effect of ground reflections is minimized during measurements but not fully suppressed, especially at low frequencies.

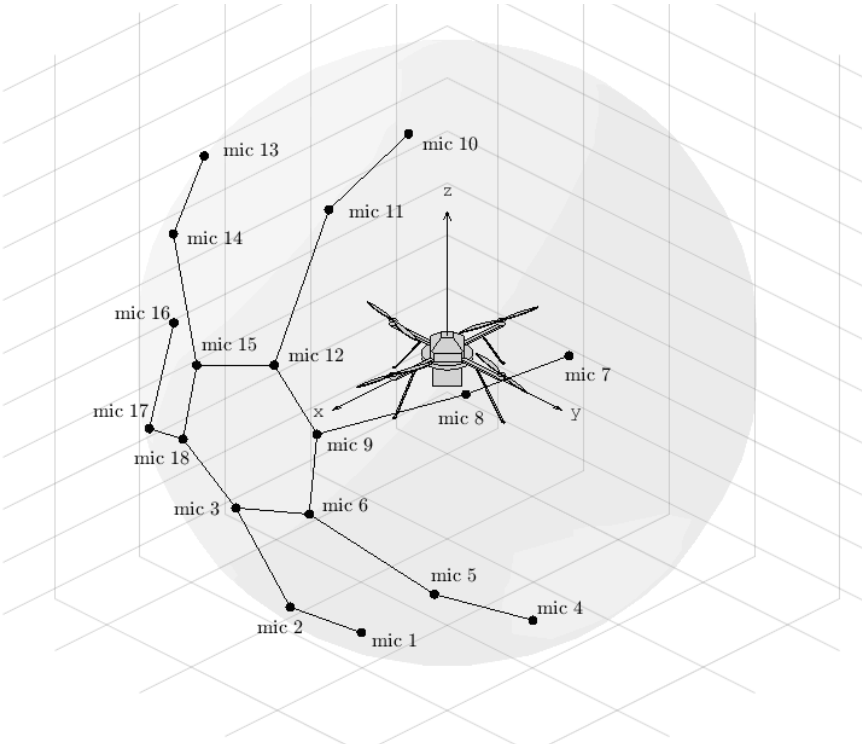


Figure 4.3: Schematic of the microphone array geometry employed in the experimental setup, with the positive x-axis direction corresponding to 0° azimuth and 0° elevation.

4.3 Results and discussion

This section presents the results on the UAV noise emission levels at different elevation angles and rotor speeds. Due to the array construction, the 0° azimuth has a higher microphone density for more elevation angles and is chosen for this analysis.

The SPL is computed considering the root-mean- squared (RMS) values of the time signals, which have been bandpass filtered between 20 Hz and 20 kHz, in particular, to discard flow-induced low-frequency pseudo-sound. For a given microphone, and therefore, elevation angle ρ, the SPL is denoted by $L_{p,\rho}$. Fig. 4.4 shows the SPL curves at 10°, −30°, −50°, and −70°, which are taken from microphones 12, 3, 2, and 1, respectively, according to Fig. 4.3. The increase in rotor speed is observed to induce an increase in SPL for all the probed elevation angles. At low rotor speeds, the $L_{p,-30°}$ is consistently the highest emission level, and the $L_{p,-70°}$ is the lowest. The $L_{p,10°}$ suffers the highest variation, being one of the highest noise emissions at low rpm and

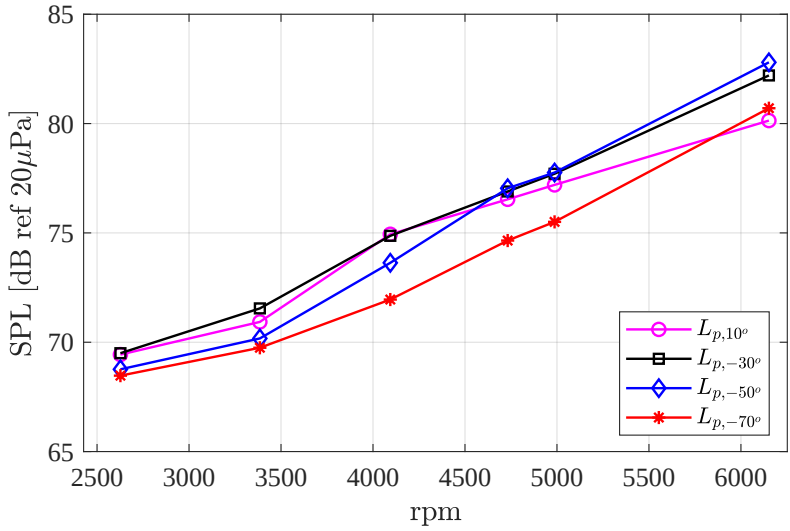
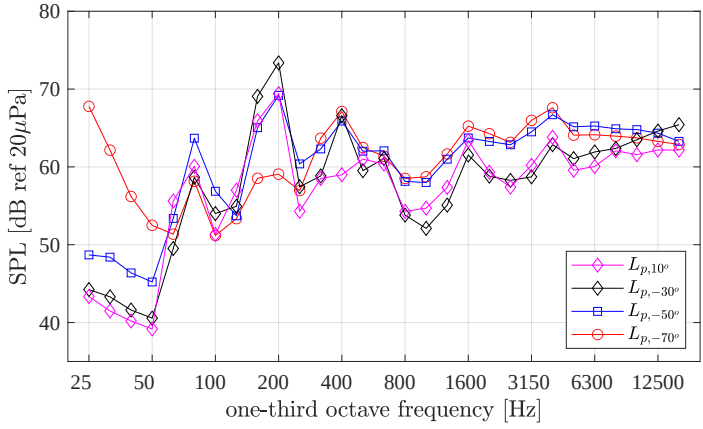


Figure 4.4: SPL for 10°, −30°, −50° and −70° emission directivity across various rotor speeds.

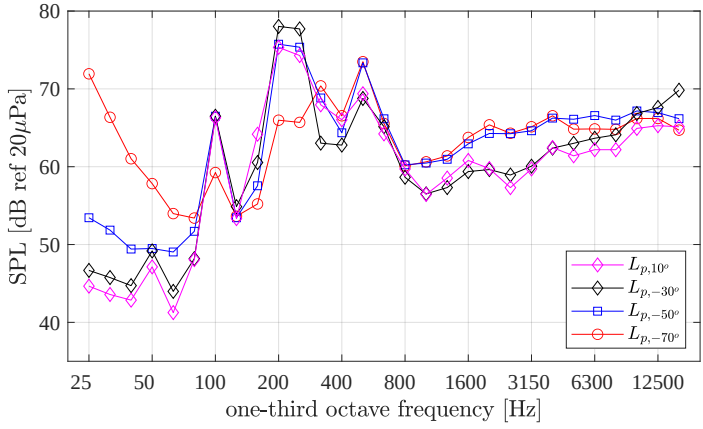
decreasing considerably at higher rpm. The decrease in level for the $L_{p,10^\circ}$ can be understood as a change in the source directivity with the throttle. Additionally, the variation in SPL among the investigated elevations at each rotor speed can be observed. For instance, at 2600 rpm, the SPL levels vary less than 1 dB across elevations, whereas at 4094 rpm, the variation can be as high as 4 dB.

In order to better assess the differences in the drone noise emission characteristics, the SPL is computed in one-third octave bands[1] for two selected rotor speeds as shown in Fig. 4.5. The one-third octave band has been chosen because it allows for easy identification of the frequency bands with higher energy when compared to the narrowband analysis and also in view of future standardization. Additionally, the one-third octave band analysis has been previously reported [207, 215, 218], which is useful for comparing the different UAVs with previous and future literature. The BPF coincides with the highest peaks between the one-third octave bands centered between 160 hertz and 200 Hz for the 4987 rpm case, and between 200 Hz and 250 Hz for the 6152 rpm case. The SPL is approximately 10 dB higher for the BPF than for the shaft rate (i.e., the first peak around 80 Hz and 100 Hz for the 4987 rpm and 6152 rpm case, respectively), with the exception of the −70° direction. This lower peak

[1] An octave band is defined as a frequency range in which the upper frequency limit corresponds to twice the lower frequency limit. A third-octave band is defined as the frequency range spanning one third of an octave, such that the upper frequency limit corresponds to $2^{1/3}$ times the lower frequency limit.



(a) 4987 rpm



(b) 6152 rpm

Figure 4.5: SPL in one-third octave bands for 0° azimuth, and 10°, −30°, −50° and −70° elevation angles.

level is in agreement with the low emission level observed in Fig. 4.4. It is interesting to notice that at 6152 rpm, the emitted levels at $-30°$ and $-50°$ are very similar (see Fig. 4.4), but their spectral content is quite different (see Fig. 4.5), especially at lower and higher frequencies. At low frequencies, below 50 Hz, the $-70°$ direction has much higher levels compared to the other directions, which can be attributed to the downward airflow across the microphone diaphragm. Similar behavior has been reported in [219]. At higher frequencies, it is possible to distinguish two trends among the four directions with higher SPL in the $10°$ and $-30°$ directions and slightly lower in the $-50°$ and $-70°$ directions. This behavior can be explained by a change in the radiation directivity with respect to the elevation angle. This trend is consistent for both rotor speeds. It is worth highlighting that the $L_{p,-70°}$ was inferred from the microphone with a wind protector without any additional correction to the signal, which can account for high-frequency attenuation at this position. In summary, the highest noise emission across all rotor speeds is found in the $-30°$ direction for the $0°$ azimuth. This observation is in agreement with other previously published works on UAVs with similar characteristics [207, 215].

4.4 Conclusion

In this chapter, an experimental methodology is proposed for measuring noise emissions from a VTOL UAV, operating at constant thrust and simulated hovering condition. Using a hemispherical microphone array setup, the methodology allows for simultaneous noise capture across various elevation and azimuth angles. The methodology is demonstrated using a quadrotor drone, and an analysis on the noise emission levels is performed for a fixed azimuth and varying elevation angles. The results obtained revealed a consistent rise in sound pressure levels (SPLs) across all observed elevation angles with increasing rotor speed. Additionally, through the analysis of the SPLs in one-third octave bands, it was observed for selected rotor speeds that while the drone noise emission levels may be similar for different elevation angles, some differences in spectral characteristics are present.

The authors note that the hemispherical microphone array employed in the proposed experimental setup has some shortcomings that need to be addressed. Firstly, the irregular distribution of microphones in the array does not provide enough sensors to validate the emissions at $90°$ azimuth, which currently limits the directivity analysis of the UAV noise. Secondly, the validity of the axis symmetry in noise emissions in the yz plane is still required to be proven (cf. [126]). In future work, besides addressing these shortcomings, the methodology can be adapted to measure other operational conditions such as takeoff and forward flight through wind tunnel and open-jet testing as long as the microphone array is not directly on the open-jet inlet. An ISO standard on noise measurements for UAVs is under development [208], and future work will aim at incorporating the requirements specified in the standard into the experimental protocol proposed in the present work. Additionally, the methodology can be used to perform in situ noise predictions and to couple with a realtime model [220] to auralize the drone fly-over in controllable and pre-defined trajectories for assessing

noise impact on the environment using emission levels, psychoacoustic metrics, and listening tests. Finally, such a platform can be of use to generate training data for drone noise detection algorithms.

Chapter 5

Speech enhancement with a UAV

This chapter is based on the publication:

Speech enhancement using ego-noise references with a microphone array embedded in an unmanned aerial vehicle

Elisa Tengan, Thomas Dietzen, Santiago Ruiz, Mansour Alkmim, João Cardenuto and Toon van Waterschoot

Published in *Proc. 2022 Int. Congr. Acoust. (ICA 2022)*, Gyeongju, South Korea, Oct. 2022, pp. 133-143.

Author contributions: ET, TD and SR jointly developed the signal processing methodology. ET, TD, MA and JC jointly designed the experimental setup and executed the experimental measurement campaign. ET implemented the algorithms and drafted the manuscript. All authors read and reviewed the final manuscript.

© 2022 The Authors. Reprinted, with permission, from:
E. Tengan, T. Dietzen, S. Ruiz, M. Alkmim, J. Cardenuto and T. van Waterschoot, "Speech enhancement using ego-noise references with a microphone array embedded in an unmanned aerial vehicle," *Proc. 2022 Int. Congr. Acoust. (ICA 2022)*, Gyeongju, South Korea, 2022, pp. 133-143.
Changes include layout, representation and minor editing.

5.1 Introduction

The use of unmanned aerial vehicles (UAVs), usually referred to as drones, has not only become more common in modern society, but also more diverse in terms of its applications and consequently, the digital signal processing solutions explored.

In situations where drones are used for cinematography, surveillance and emergency search and rescue operations, the potential of recording and processing audio signals, as opposed to limiting the use of UAVs to image and video capture only, has become more evident and researched over the past few years [221,222]. A fundamental problem in recording audio with a UAV, however, is the high level of ego-noise generated by its rotors [111, 223], which interferes with the target source signal and consequently reduces sound quality. In order to overcome this issue, noise reduction and speech enhancement techniques can be employed [63].

In the current state of the art, noise reduction and speech enhancement frameworks based on the standard multichannel Wiener filter (MWF), computed as a function of the speech source steering vector, or on blind source separation (BSS) methods, have already been used and tested on UAV setups [112, 222]. However, these approaches require the speech source location to be presumably known, which can be a limiting factor in more practical scenarios. An alternative formulation of the standard MWF allows the filter coefficients to be computed as a function of the speech correlation matrix instead [65], which is in turn estimated by using speech activity information obtained with a voice activity detector (VAD) or with speech presence probability (SPP) estimates [75]. When considering this formulation, an extension of the standard MWF can be realized when prior knowledge on the noise is available from specific channels of the microphone array and used to improve robustness in the estimation of the speech correlation matrix, resulting in the prior knowledge multichannel Wiener filter (PK-MWF) [83].

In this paper, a method is proposed for performing speech enhancement using ego-noise references with a microphone array embedded in a UAV, while assuming that the source location, and consequently, the source steering vector with respect to the array configuration, is unknown. The ego-noise reference signals are captured with microphones located near the UAV's propellers and used to constrain the estimation of the speech correlation matrix necessary for computing the prior knowledge multichannel Wiener filter (PK-MWF). Speech presence probability (SPP) information is estimated in order to identify speech activity in different time-frequency points. Such SPP estimation can either be obtained from one of the embedded array's microphones or from an external microphone also used for recording the auditory scene and providing a performance benchmark. Experimental measurements, made available online [123], were performed in a semi-anechoic chamber for testing the implementation of the standard MWF and the PK-MWF. It will be shown that the use of PK-MWF achieves higher levels of improvement than the standard MWF in terms of short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ), while the SNR improvement is similar.

The paper is organized as follows. In section 5.2, the signal model is defined. In section 5.3, the proposed method is explained and its formulation is compared with the standard multichannel Wiener filter (MWF). In section 5.4, the experimental setup used for the recordings and the audio processing framework are described. In section 5.5, the results obtained are discussed and finally, in section 5.6, a conclusion is presented with a summary and final remarks on the work accomplished.

5.2 Signal model

The signal in the short-time Fourier transform (STFT) domain captured by a single microphone with index m is defined as

$$y_m(\omega_k, l) = a_m(\omega_k, l)\, s(\omega_k, l) + n_m(\omega_k, l), \tag{5.1}$$

where y_m is the resulting signal, a_m is the corresponding transfer function from the target source to the microphone, s is the speech source signal, and n_m is the noise component. The terms ω_k and l correspond to the discrete frequency, indexed by k, and the observation frame index, respectively. In the following equations, we consider processing the different frames and frequency bins independently, and their corresponding indices will be omitted for brevity.

Considering a microphone array with M elements, a signal vector is obtained by stacking all microphone signal equations, resulting in

$$\mathbf{y} = \mathbf{s} + \mathbf{n} \tag{5.2}$$

$$= \mathbf{a}s + \mathbf{n}, \tag{5.3}$$

with $\{\mathbf{y}, \mathbf{s}, \mathbf{a}, \mathbf{n}\} \in \mathbb{C}^M$, and $s \in \mathbb{C}$. Assuming that the desired target source signal and the noise component are uncorrelated, the correlation matrices adhere to

$$\mathbf{R_{yy}} = \mathbb{E}\left\{\mathbf{y}\mathbf{y}^{\mathrm{H}}\right\} \tag{5.4}$$

$$= \mathbb{E}\left\{\mathbf{s}\mathbf{s}^{\mathrm{H}}\right\} + \mathbb{E}\left\{\mathbf{n}\mathbf{n}^{\mathrm{H}}\right\} \tag{5.5}$$

$$= \mathbf{R_{ss}} + \mathbf{R_{nn}}, \tag{5.6}$$

where $\mathbb{E}\{.\}$ denotes the expectation operator, $(.)^{\mathrm{H}}$ denotes the conjugate transpose, $\mathbf{R_{yy}}$ is the microphone signal correlation matrix, $\mathbf{R_{ss}}$ and $\mathbf{R_{nn}}$ are the speech signal and noise correlation matrices, respectively. The speech correlation matrix corresponds to $\mathbf{R_{ss}} = \mathbf{a}\mathbf{a}^{\mathrm{H}}\mathbb{E}\left\{s^2\right\}$, and hence it is rank-1 under the common assumption that $\mathbf{a}$ is deterministic.

5.3 Proposed method

We propose to use the PK-MWF with prior knowledge obtained from microphones mounted below the rotors of the UAV. In section 5.3.1, we introduce the general formulation of the multi-channel Wiener filter. In section 5.3.2, we outline the standard MWF realization, and in section 5.3.3, we describe the PK-MWF realization and its application in the context of a UAV.

5.3.1 Formulation of the multichannel Wiener filter (MWF)

Let a target speech reference signal d be defined as

$$d = \mathbf{e}_d^\top \mathbf{s}, \tag{5.7}$$

with $\mathbf{e}_d = [1, 0, \ldots, 0]^\top$ being a vector selecting the desired source signal component in the microphone array's first channel, which is here without loss of generality considered as the reference channel, and $(.)^\top$ denoting the transpose.

In the formulation of the multichannel Wiener filter (MWF), it is aimed to estimate d as a linear combination of all microphone signals in $\mathbf{y}$ by minimizing the following mean squared error (MSE) criterion and obtaining the filter weights $\bar{\mathbf{w}}$ as [65]

$$\bar{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbb{E}\left\{ |d - \mathbf{w}^{\mathrm{H}}\mathbf{y}|^2 \right\}. \tag{5.8}$$

If the microphone signal correlation matrix $\mathbf{R_{yy}}$ has full rank, the optimal solution to (5.8) is [65]

$$\bar{\mathbf{w}} = \mathbf{R_{yy}^{-1}}\mathbf{R_{ss}}\mathbf{e}_d. \tag{5.9}$$

Since the speech-only correlation matrix $\mathbf{R_{ss}}$ cannot be directly observed from the microphone signals due to the presence of noise, its estimate can be based on the analysis of the speech activity's on-off behavior in the time-frequency domain. While assuming short-term stationarity of $\mathbf{y}$, the estimation of the speech-plus-noise and noise-only correlation matrices ($\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$, respectively) can be performed instead and used to estimate $\hat{\mathbf{R}}_{\mathbf{ss}}$.

5.3.2 Standard multichannel Wiener filter (MWF) realization

In the standard multichannel Wiener filter, the speech-only correlation matrix is estimated by solving the following optimization problem [224]:

$$\hat{\mathbf{R}}_{\mathbf{ss}} = \underset{\substack{\mathrm{rank}(\mathbf{R_{ss}})=1 \\ \mathbf{R_{ss}} \succeq 0}}{\arg\min} \left\| \hat{\mathbf{R}}_{\mathbf{nn}}^{-\frac{1}{2}} \left(\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}} - \mathbf{R_{ss}} \right) \hat{\mathbf{R}}_{\mathbf{nn}}^{-\frac{\mathrm{H}}{2}} \right\|_{\mathrm{F}}^2, \tag{5.10}$$

where $\|.\|_{\mathrm{F}}$ denotes the Frobenius norm and $\hat{\mathbf{R}}_{\mathbf{ss}}$ is constrained to be a rank-1 and positive semi-definite matrix. The solution to (5.10) is based on the generalized eigenvalue decomposition (GEVD) of the matrix pencil $\{\hat{\mathbf{R}}_{\mathbf{yy}}, \hat{\mathbf{R}}_{\mathbf{nn}}\}$ [224]:

$$\hat{\mathbf{R}}_{\mathbf{yy}} = \hat{\mathbf{Q}}\hat{\mathbf{\Sigma}}_{\mathbf{yy}}\hat{\mathbf{Q}}^{\mathrm{H}}, \tag{5.11}$$

$$\hat{\mathbf{R}}_{\mathbf{nn}} = \hat{\mathbf{Q}}\hat{\mathbf{\Sigma}}_{\mathbf{nn}}\hat{\mathbf{Q}}^{\mathrm{H}}, \tag{5.12}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{yy}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{nn}}$ are diagonal matrices containing the generalized eigenvalues $\hat{\sigma}_{y_i}$ and $\hat{\sigma}_{n_i}$ of $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$, respectively, and $\hat{\mathbf{Q}}$ is an invertible matrix containing the generalized eigenvectors in its columns. The generalized eigenvalues and eigenvectors are assumed to be sorted in descending order of $\hat{\sigma}_{y_i}$ and $\hat{\sigma}_{n_i}$.

Using (5.6), the estimate of $\hat{\mathbf{R}}_{\mathbf{ss}}$ can then be obtained as

$$\hat{\mathbf{R}}_{\mathbf{ss}} = \hat{\mathbf{Q}}\,\mathrm{diag}\{\hat{\sigma}_{y_1} - \hat{\sigma}_{n_1}, 0, \ldots, 0\}\hat{\mathbf{Q}}^{\mathrm{H}}, \tag{5.13}$$

with diag{.} building a diagonal matrix and $\hat{\sigma}_{y_1}$ and $\hat{\sigma}_{n_1}$ being the top-left elements of $\hat{\mathbf{\Sigma}}_{\mathbf{yy}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{nn}}$, respectively, yielding the largest ratio between eigenvalues $\hat{\sigma}_{y_i}/\hat{\sigma}_{n_i}$. Finally, by substituting (5.13) and (5.11) into (5.9), the standard multichannel Wiener filter can be formulated as [224]

$$\hat{\mathbf{w}}_{\mathrm{MWF}} = \hat{\mathbf{Q}}^{-\mathrm{H}}\mathrm{diag}\left\{1 - \frac{\hat{\sigma}_{n1}}{\hat{\sigma}_{y1}}, 0, \ldots, 0\right\}\hat{\mathbf{Q}}^{\mathrm{H}}\mathbf{e}_d. \tag{5.14}$$

5.3.3 Prior knowledge multichannel Wiener filter (PK-MWF) realization

In the standard multichannel Wiener filter formulation, the spatial configuration of the microphone array used with respect to the noise sources is not explicitly exploited. However, for a practical setup such as the microphone array embedded in a UAV considered in this study, the prior knowledge obtained from array elements sufficiently close to the main noise sources interfering in the signals, i.e. the device's rotors, can be used in an attempt to provide a more robust estimation of $\mathbf{R}_{\mathbf{ss}}$.

We define $M_{\mathrm{S+N}}$ as the number of array elements that are assumed to contain speech and noise, and M_{N} as the number of array elements assumed to only contain noise, or have a sufficiently low signal-to-noise ratio (SNR) such that the speech component is negligible. In a UAV setup, such configuration is here assumed to be attained if each of the M_{N} microphones are placed in close proximity to one of the vehicle's propellers. The total number of microphones used are then described as $M = M_{\mathrm{S+N}} + M_{\mathrm{N}}$.

Let the identity and all-zero matrix be denoted by $\mathbf{I}$ and $\mathbf{0}$, respectively, where we indicate their dimensions by a subscript. We define the selection matrix $\mathbf{H}$ and the

blocking matrix $\mathbf{B}$ as

$$\mathbf{H} = \begin{bmatrix} \mathbf{I}_{M_{\mathrm{S+N}}} \\ \mathbf{0}_{M_{\mathrm{N}} \times M_{\mathrm{S+N}}} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \mathbf{0}_{M_{\mathrm{S+N}} \times M_{\mathrm{N}}} \\ \mathbf{I}_{M_{\mathrm{N}}} \end{bmatrix}. \tag{5.15}$$

Thus, the selection matrix $\mathbf{H}$ and blocking matrix $\mathbf{B}$ have dimensions $M \times M_{\mathrm{S+N}}$ and $M \times M_{\mathrm{N}}$, respectively. In the PK-MWF, the following cost function is then minimized in order to estimate $\hat{\mathbf{R}}_{\mathbf{ss}}$ [83]:

$$\hat{\mathbf{R}}_{\mathbf{ss}} = \underset{\substack{\mathrm{rank}(\mathbf{R}_{\mathbf{ss}})=1 \\ \mathbf{B}^{\mathrm{H}}\mathbf{R}_{\mathbf{ss}}\mathbf{B}=\mathbf{0} \\ \mathbf{R}_{\mathbf{ss}} \succeq 0}}{\arg\min} \left\| \hat{\mathbf{R}}_{\mathbf{nn}}^{-\frac{1}{2}} \left(\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}} - \mathbf{R}_{\mathbf{ss}} \right) \hat{\mathbf{R}}_{\mathbf{nn}}^{-\frac{\mathrm{H}}{2}} \right\|_{\mathrm{F}}^{2}. \tag{5.16}$$

The matrix $\hat{\mathbf{R}}_{\mathbf{ss}}$ to be estimated now, which is the counterpart to (5.10) in the standard MWF, is not only constrained to be rank-1 and positive semi-definite, but also to have its column and row spaces lying in the column space of $\mathbf{H}$, such that $\mathbf{B}^{\mathrm{H}}\mathbf{R}_{\mathbf{ss}}\mathbf{B} = \mathbf{0}$. The solution to (5.16) (proof omitted) is obtained by firstly applying a linearly-constrained minimum variance (LCMV) beamformer $\mathbf{C}$ to the vector $\mathbf{y}$, defined by the following criterion [83]:

$$\mathbf{C} = \underset{\mathbf{C}}{\arg\min} \quad \mathrm{trace}\left\{ \mathbf{C}^{\mathrm{H}}\mathbf{R}_{\mathbf{nn}}\mathbf{C} \right\} \tag{5.17}$$

$$\text{s.t.} \quad \mathbf{H}^{\mathrm{H}}\mathbf{C} = \mathbf{I}_{M_{\mathrm{S+N}}}, \tag{5.18}$$

which has its solution based on a generalized sidelobe canceler (GSC) [225]:

$$\hat{\mathbf{C}} = \mathbf{H} - \mathbf{B}\hat{\mathbf{F}}, \tag{5.19}$$

$$\hat{\mathbf{F}} = \left(\mathbf{B}^{\mathrm{H}}\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{B} \right)^{-1} \mathbf{B}^{\mathrm{H}}\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{H}. \tag{5.20}$$

By applying $\hat{\mathbf{C}}$ to $\mathbf{y}$, we obtain a vector with reduced dimension $\mathbf{y}^{\mathrm{red}} = \hat{\mathbf{C}}^{\mathrm{H}}\mathbf{y}$. Similarly, we can obtain the reduced dimension correlation matrices $\hat{\mathbf{R}}_{\mathbf{yy}}^{\mathrm{red}} = \hat{\mathbf{C}}^{\mathrm{H}}\hat{\mathbf{R}}_{\mathbf{yy}}\hat{\mathbf{C}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}^{\mathrm{red}} = \hat{\mathbf{C}}^{\mathrm{H}}\hat{\mathbf{R}}_{\mathbf{nn}}\hat{\mathbf{C}}$. Then, a generalized eigenvalue decomposition (GEVD) of the reduced matrix pencil $\left\{ \hat{\mathbf{R}}_{\mathbf{yy}}^{\mathrm{red}}, \hat{\mathbf{R}}_{\mathbf{nn}}^{\mathrm{red}} \right\}$ is performed as in [83]

$$\hat{\mathbf{R}}_{\mathbf{yy}}^{\mathrm{red}} = \hat{\mathbf{Q}}^{\mathrm{red}}\hat{\mathbf{\Sigma}}_{\mathbf{yy}}^{\mathrm{red}}(\hat{\mathbf{Q}}^{\mathrm{red}})^{\mathrm{H}}, \tag{5.21}$$

$$\hat{\mathbf{R}}_{\mathbf{nn}}^{\mathrm{red}} = \hat{\mathbf{Q}}^{\mathrm{red}}\hat{\mathbf{\Sigma}}_{\mathbf{nn}}^{\mathrm{red}}(\hat{\mathbf{Q}}^{\mathrm{red}})^{\mathrm{H}}, \tag{5.22}$$

which is the PK-MWF counterpart to (5.11)-(5.12) in the standard MWF. Again, we assume the generalized eigenvalues and eigenvectors to be sorted in descending order of $\hat{\sigma}_{y_i}^{\mathrm{red}}$ and $\hat{\sigma}_{n_i}^{\mathrm{red}}$. The matrix $\hat{\mathbf{R}}_{\mathbf{ss}}$ can then be expressed as [83]

$$\hat{\mathbf{R}}_{\mathbf{ss}} = \mathbf{H}\hat{\mathbf{Q}}^{\mathrm{red}}\ \mathrm{diag}\left\{ \hat{\sigma}_{y1}^{\mathrm{red}} - \hat{\sigma}_{n1}^{\mathrm{red}}, 0, ..., 0 \right\} (\hat{\mathbf{Q}}^{\mathrm{red}})^{\mathrm{H}}\mathbf{H}^{\mathrm{H}}, \tag{5.23}$$

which corresponds to the PK-MWF counterpart to (5.13) in the standard MWF. Finally, by substituting (5.23) and (5.21) into (5.9), the estimated prior knowledge multichannel Wiener filter (PK-MWF) coefficients are given by

$$\hat{\mathbf{w}}_{\text{PK-MWF}} = \hat{\mathbf{C}}(\hat{\mathbf{Q}}^{\text{red}})^{-\text{H}}\text{diag}\left\{1 - \frac{\hat{\sigma}_{n1}^{\text{red}}}{\hat{\sigma}_{y1}^{\text{red}}}, 0, \ldots, 0\right\}(\hat{\mathbf{Q}}^{\text{red}})^{\text{H}}\mathbf{H}^{\text{H}}\mathbf{e}_d. \tag{5.24}$$

5.4 Evaluation

In order to compare the performance of both multichannel filtering methods specified in section 5.3, experimental tests were carried out in a semi-anechoic chamber at Siemens Digital Industries Software, in Leuven, Belgium. In section 5.4.1, the measurement setup is detailed, and in section 5.4.2, the processing framework and the different cases studied are described.

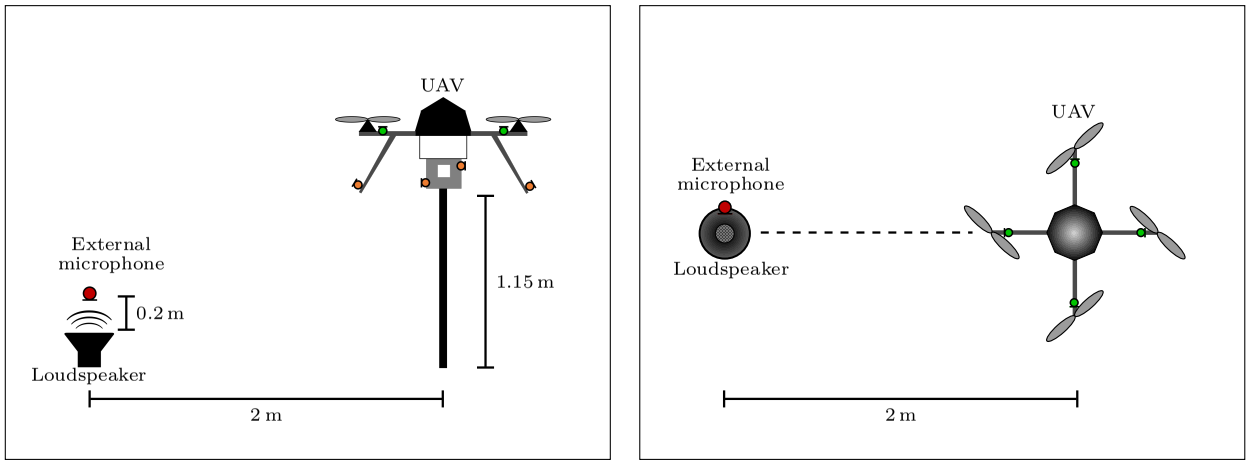
5.4.1 Measurement setup

In the measurement setup considered for performance evaluation, a MikroKopter MK EASY Quadro V3 (HiSystems GmbH) quad-rotor UAV is mounted on a support stand such that the bottom of its custom-made frame is 1.15 m above the floor. The UAV is equipped with a 16-element array composed of electret condenser microphones, a sound card and a minicomputer for performing audio recordings. A loudspeaker is placed 2 m away from the base of the stand and close to ground level, in order to simulate a scenario of a speech source being present below the UAV's line-of-sight. An external microphone is placed 0.2 m above the loudspeaker for recording the reference speech signal and allowing a performance comparison in the processing stage. A sketch representation is depicted in Fig. 5.1, and a photo of the actual measurement setup inside the semi-anechoic chamber is presented in Fig. 5.2.

With the UAV's throttle level fixed to a nearly constant value, which was possible due to the remote control's throttle joystick not being spring-loaded, the setup conditions could be considered an approximation of a hovering state for the UAV [226]. A male speech signal from the VCTK corpus [227] was played from the loudspeaker and recorded both by the external reference microphone and the drone's microphone array. The recordings, available online [123], were repeated for three different levels of throttle, and therefore, three different propeller rotational speeds, measured at run time with a laser probe placed under one of the propeller blades.

5.4.2 Processing

The measurements were processed offline on Matlab by, firstly, downsampling the recorded signals from 44.1 kHz to 16 kHz. Then, a 512-point square-root of Hann observation window with 50% overlap is employed for computing the FFT for the



(a) Side view

(b) Top view

Figure 5.1: Sketches of experimental setup

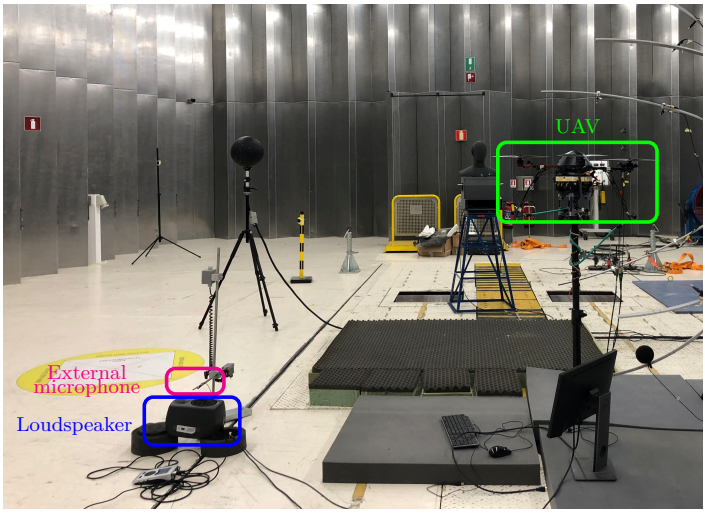


Figure 5.2: Experimental setup in semi-anechoic chamber

signal frames from each microphone. The correlation matrices $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$ are estimated for each frequency bin based on the speech presence probability (SPP) [71] value of all frames being processed. Such SPP can be either estimated from one of the channels from the embedded microphone array (denoted as iSPP) or alternatively from the reference external microphone (denoted as xSPP), in order to provide a performance benchmark in the succeeding analysis of results obtained. We define $\beta(\omega_k, l)$ as a speech activity indicator, which is computed as

$$\beta(\omega_k, l) = \begin{cases} 1 \text{ (speech active)}, & \text{if SPP}(\omega_k, l) \geq 0.5 \\ 0 \text{ (speech inactive)}, & \text{otherwise.} \end{cases} \tag{5.25}$$

This parameter is then used to estimate the desired correlation matrices as

$$\hat{\mathbf{R}}_{\mathbf{yy}}(\omega_k) = \frac{1}{L_{\text{ON}}(\omega_k)} \sum_{l=1}^{L} \mathbf{y}(\omega_k, l)\mathbf{y}^{\text{H}}(\omega_k, l)\beta(\omega_k, l), \tag{5.26}$$

$$\hat{\mathbf{R}}_{\mathbf{nn}}(\omega_k) = \frac{1}{L_{\text{OFF}}(\omega_k)} \sum_{l=1}^{L} \mathbf{y}(\omega_k, l)\mathbf{y}^{\text{H}}(\omega_k, l)(1 - \beta(\omega_k, l)), \tag{5.27}$$

where L denotes the total number of frames processed, and $L_{\text{ON}}(\omega_k)$ and $L_{\text{OFF}}(\omega_k)$ correspond to the total number of frames for frequency ω_k where, based on the value of $\beta(\omega_k, l)$ being 1 or 0, the desired speech component is assumed to be active or inactive, respectively.

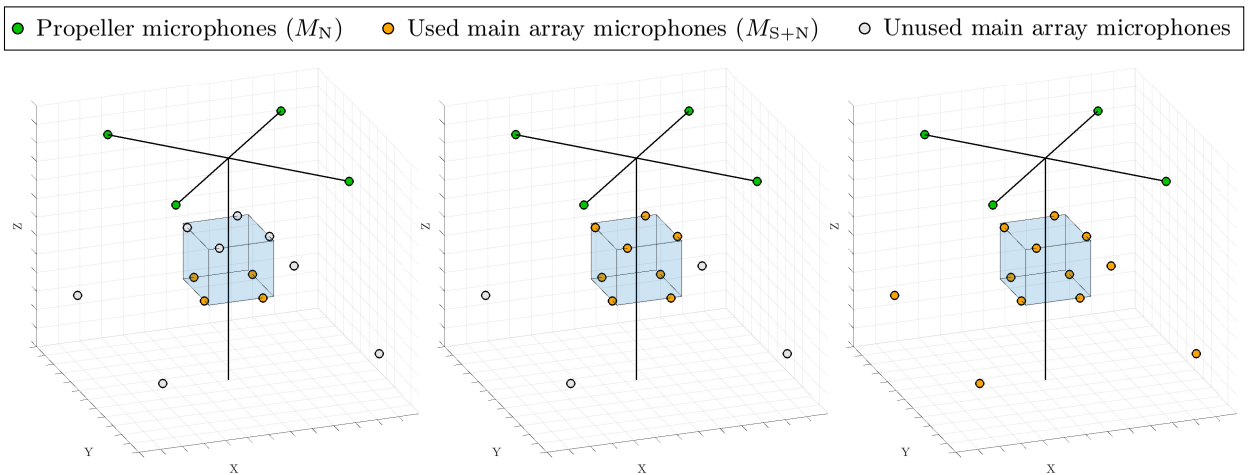
In order to observe possible variations in performance according to different microphone array configurations, the MWF and PK-MWF filters are implemented considering three different numbers of channels used from the embedded microphone array, as illustrated in Fig. 5.3. In the PK-MWF implementation, the number of microphones from the main array used, denoted by $M_{\mathrm{S+N}}$, corresponds to the number of channels employed in the selection matrix $\mathbf{H}$, whereas the four propeller microphones, placed below each pair of blades and denoted by M_{N}, are always used as noise references in the blocking matrix $\mathbf{B}$. In the case of MWF, however, we consider two implementations where the propeller microphones M_{N} are used in its formulation or not, in addition to the channels used from the main array $M_{\mathrm{S+N}}$.

The performance of the different methods implemented is evaluated in terms of SNR improvement, as well as of two other objective measures, namely the short-time objective intelligibility (STOI) [228] and the perceptual evaluation of speech quality (PESQ) [229].

5.5 Results

The resulting SNR improvement obtained at the channel where the reference signal estimate $\hat{d} = \hat{\mathbf{w}}^{\mathrm{H}}\mathbf{y}$ is computed is presented in Fig. 5.4. Different numbers of microphones in the proposed filtering formulations and the SPP estimates were considered. The bottom and top horizontal axes indicate the original SNRs considered and the corresponding average rotational speed of the rotors in revolutions/min (rpm), respectively. We can observe that, with the SPP estimate from the reference external signal (xSPP), the SNR improvement is greater than when the SPP estimate from the embedded array's own reference channel (iSPP) is used, as the external microphone placed near the target source provides a signal with greater SNR itself. Therefore, the detection of speech activity is more reliable, resulting in more accurate estimates of the necessary correlations matrices considered here. Since having an external microphone near the target speaker may not be realistic in all kinds of scenarios, the use of the external SPP estimate can be here seen as a way to provide a performance benchmark for the proposed methods, with which it is possible to compare and evaluate the improvements obtained and its possible limitations. In this case, it is observed that even with a poorer SPP estimate, all methods are still capable of improving the SNR of the target reference channel.

In terms of the number of microphones used for performing the multichannel filtering, it is observed that performance is improved when using more microphone signals, although the differences are more visible when going from $M_{\mathrm{S+N}} = 4$ to $M_{\mathrm{S+N}} = 8$ than when going from $M_{\mathrm{S+N}} = 8$ to $M_{\mathrm{S+N}} = 12$. This is possibly due to performance saturation related to signal model or SPP errors that could not be improved with the increase in number of microphones used, as well as to the microphone placement of the last four channels included being aligned with the rotors' rotational axes, which is considered to be where the noise level is the strongest [111] and therefore, not providing as much additional information on the target speech signal as the previously included



(a) $M_{\mathrm{S+N}} = 4$, $M_{\mathrm{N}} = 4$ (b) $M_{\mathrm{S+N}} = 8$, $M_{\mathrm{N}} = 4$ (c) $M_{\mathrm{S+N}} = 12$, $M_{\mathrm{N}} = 4$

Figure 5.3: Illustration of different microphone array configurations considered in the formulation of MWF and PK-MWF.

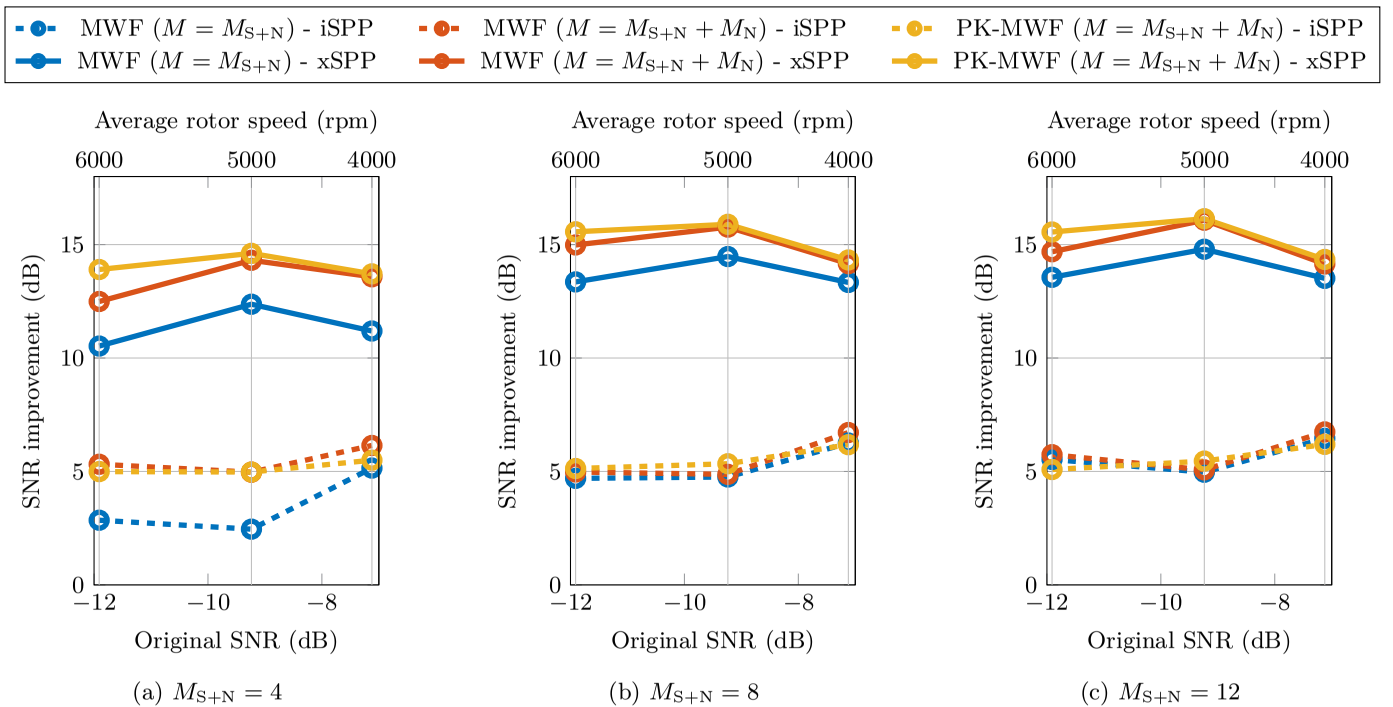
microphones. We can also observe that the performance of PK-MWF in terms of SNR improvement is similar to the one of MWF while employing the same number of microphones ($M = M_{\mathrm{S+N}} + M_{\mathrm{N}}$). However, the matrix reductions performed in the PK-MWF might favor its usage in terms of computational complexity.

The improvement in speech intelligibility measured by STOI is presented in Fig. 5.5. It can be observed that the performance of PK-MWF with respect to this perceptual measure is particularly better than both formulations of standard MWF considered when employing the SPP estimate from the embedded array's own reference signal (iSPP), indicating the advantage of using the ego-noise reference signals as proposed here for improving robustness to SPP errors which affect the estimation of $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$. In addition to having a similar behavior in performance improvement when including more channels as the one seen in terms of SNR, we can also observe that the STOI improvement rate increases with the original values from the array's reference signal, which corresponds to a decrease in the average rotor speed.

Finally, the improvement in speech quality measured by PESQ is presented in Fig. 5.6, which indicates a better performance of PK-MWF compared to the standard MWF filters when using either one of the SPP estimates considered. As opposed to the STOI improvement depicted in Fig. 5.5, the performance improvement rate in this case decreases with the original PESQ score of the reference signal considered, which is in turn associated to an increase in the average rotational speed of the rotors.

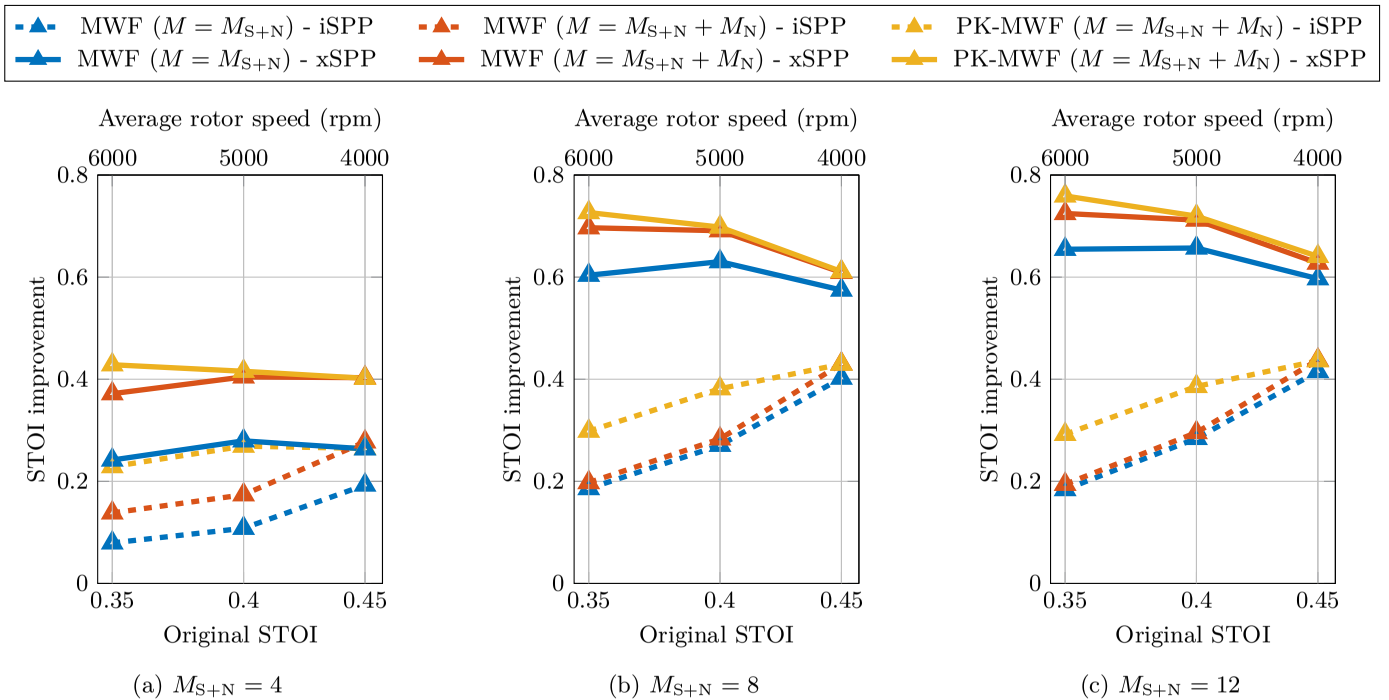
5.6 Conclusion

In this paper, a method for speech enhancement using ego-noise references with a microphone array embedded in an unmanned aerial vehicle was proposed. The ego-noise reference signals captured by microphones placed near the UAV's rotors were used as prior knowledge in the PK-MWF for estimating the necessary speech correlation matrix in a constrained optimization problem. Speech presence probability (SPP) was estimated from both an external reference microphone and one of the embedded array's channels in order to detect speech activity. Results obtained from experimental recordings showed that, especially in a more realistic scenario where the SPP is estimated from one of the UAV's own microphones, the PK-MWF implementation provided greater improvement in speech intelligibility and quality when compared to the use of standard MWF, with and without the propeller microphones being included in its formulation. While the SNR improvement is similar for PK-MWF and the standard MWF using the propeller microphones, it can be argued that employing the PK-MWF implementation might be advantageous in terms of computational complexity due to the performed matrix reductions when considering real-time processing applications. Future work includes expanding the measurement campaign to consider moving UAVs, updating the correlation matrices over different time frames, and reformulating the noise signal model to consider more particular characteristics of UAVs' ego-noise.



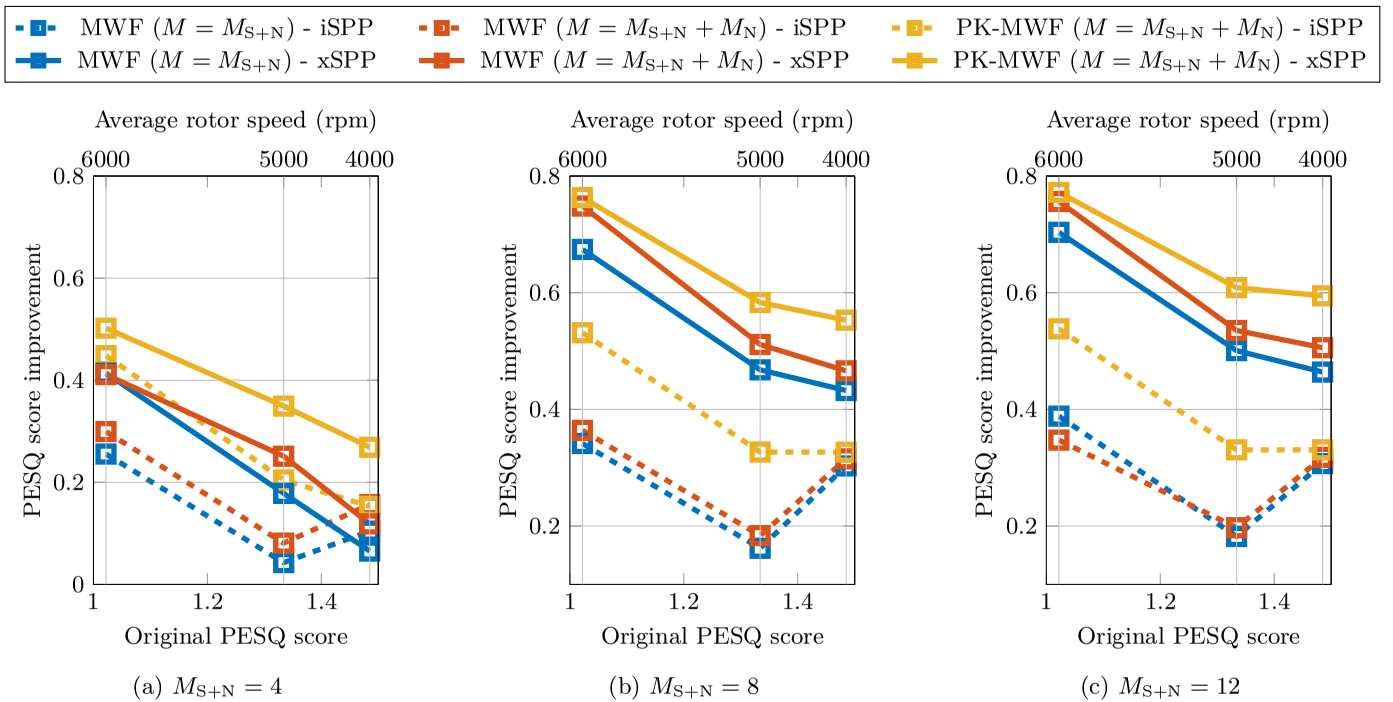
(a) $M_{S+N} = 4$ (b) $M_{S+N} = 8$ (c) $M_{S+N} = 12$

Figure 5.4: SNR improvement in function of original SNRs associated to different average rotor speeds.



(a) $M_{S+N} = 4$ (b) $M_{S+N} = 8$ (c) $M_{S+N} = 12$

Figure 5.5: STOI improvement in function of original STOI values associated to different average rotor speeds.



(a) $M_{S+N} = 4$ (b) $M_{S+N} = 8$ (c) $M_{S+N} = 12$

Figure 5.6: PESQ improvement in function of original PESQ scores associated with different average rotor speeds.

Chapter 6

DOA-informed speech enhancement with a UAV

6.1 Introduction

As presented in **Ch. 5**, the task of speech enhancement using audio recordings obtained from a microphone array embedded in an unmanned aerial vehicle (UAV) presents significant challenges. When employing the ego-noise references to formulate the prior-knowledge multichannel Wiener filter (PK-MWF), experimental results highlighted the influence of the speech activity estimation on the speech enhancement performance. However, relying on an external microphone as a reference for obtaining an accurate speech activity estimate is often impractical in many applications. Therefore, this chapter focuses on the development of a strategy to improve the performance of speech enhancement with a UAV.

Based on the relevance of a priori spatial information in many speech enhancement approaches [80, 81, 84], the study presented in this chapter aims to incorporate the target speech source's direction of arrival (DOA) into the PK-MWF implementation, in order to overcome limitations posed by unavailable or unreliable speech activity estimates.

Since, in this context, a target speech source is not expected to be located on the same horizontal plane as the UAV, the target source's DOA is estimated in terms of both azimuth and elevation, using the method presented in **Ch. 3**, which is based on the solution of a group-sparse regularized optimization problem modeling a broadband steered response power (SRP) as a linear combination of location-specific PSDs. Considering the specific noise conditions that the recordings obtained with a microphone array embedded in a UAV would face in most practical acoustic scenarios, the computation of the observed SRP map is here modified by explicitly including a weighting factor related to the UAV ego-noise's spectral characteristics.

Following the DOA estimation, the corresponding direct-path relative transfer function (RTF) estimate is integrated into the implementation of the PK-MWF filter from **Ch. 5**. Previously, the multi-channel Wiener filter (MWF) stage of PK-MWF, following the generalized sidelobe canceller (GSC), relied on a generalized eigenvalue decomposition (GEVD) of reduced correlation matrices, which in turn were estimated as a function of speech presence probability (SPP). However, in this context, the MWF stage is restructured using its well-established alternative formulation, comprising a minimum variance distortionless response (MVDR) beamformer, followed by a single-channel Wiener filter estimate [8, 63]. This reformulation allows for incorporating the direct-path RTF corresponding to the target source's estimated DOA into the proposed speech enhancement framework.

We evaluate the proposed alternative to the previous PK-MWF approach of **Ch. 5** in terms of the same objective measures considered (i.e., signal-to-noise ratio (SNR), short-time objective intelligibility (STOI) [228] and perceptual evaluation of speech quality (PESQ) [229]), and results indicate that employing the DOA-based implementation of the PK-MWF filter allows for greater improvement in SNR of the filtered signal, when compared to the result obtained when employing the PK-MWF with SPP estimated from one of the embedded microphones. Conversely, the perceptual

improvement, in terms of STOI and PESQ, is compromised. Furthermore, it is shown that when processing different frequency bands independently, as is commonly done in spatial filtering methods, it is possible to combine the DOA and SPP-based implementations of the PK-MWF. Results indicate that such combination of the two approaches can improve performance in terms of PESQ, at the expense of lower SNR improvement when compared to the results obtained while solely employing the DOA-based implementation.

The remainder of this chapter is organized as follows. In Sec. 6.2, the use of the DOA estimation method proposed in **Ch. 3** is detailed, while considering the specific scenario of recordings made with a microphone array embedded in a UAV. In Sec. 6.3, the incorporation of the estimated target speech source DOA into the PK-MWF realization is explained. In Sec. 6.4, the evaluation of the modified PK-MWF is presented, and Sec. 6.5 concludes the chapter.

6.2 Direction-of-arrival estimation using a microphone array embedded in a UAV

6.2.1 Broadband SRP in terms of azimuth and elevation

In **Ch. 3**, it is assumed that broadband sources are located in the far field, in the same horizontal plane as a planar microphone array, such that their location is defined solely in terms of azimuth. In the context of UAV applications, however, this assumption does not usually hold. Hence, we revisit the definition of the time difference of arrival (TDOA) in (3.1) and alternatively express it as a function of both azimuth and elevation, denoted γ and ρ, respectively:

$$\Delta t_{m,m'}(\gamma,\rho) = (\mathbf{p}_m - \mathbf{p}_{m'})^\top \mathbf{r}_{\gamma,\rho}/c, \tag{6.1}$$

where, according to the coordinate system from **Ch. 4** (cf. Fig. 4.2), the incident direction vector $\mathbf{r}_{\gamma,\rho}$ is given by

$$\mathbf{r}_{\gamma,\rho} = -[\cos\gamma\cos\rho, \sin\gamma\cos\rho, \sin\rho]^\top. \tag{6.2}$$

One may note that the broadband SRP formulation in (3.9) can then be readily extended for DOAs expressed in terms of both azimuth γ and elevation ρ as

$$z(\gamma,\rho) = 2\sum_{p=1}^{P}\sum_{k=1}^{K} \Re\{\psi_p(\omega_k) e^{j\omega_k \Delta t_p(\gamma,\rho)}\}, \tag{6.3}$$

with the definitions of P, p, K, k and $\psi_p(\omega_k)$ being equal to those in **Ch. 3**.

6.2.2 Weighted broadband SRP based on the spectral characteristics of UAV ego-noise

As an attempt to pre-emptively improve the robustness of the DOA estimation method presented in **Ch. 3**, under the specific acoustic conditions that UAV recordings would encounter, we propose to account for the spectral characteristics of the UAV ego-noise in the computation of the broadband SRP map.

Assuming that the UAV ego-noise is stationary and that noise-only microphone recordings are available, let $\hat{\sigma}_{n,\mathrm{dB}}(\omega_k)$ denote the corresponding estimated noise PSD, in dB, at frequency ω_k. Additionally, let $\Delta\hat{\sigma}_{n,\mathrm{dB}}$ denote the difference between the maximum and minimum estimated PSD values, given by

$$\Delta\hat{\sigma}_{n,\mathrm{dB}} = \max_k \hat{\sigma}_{n,\mathrm{dB}}(\omega_k) - \min_k \hat{\sigma}_{n,\mathrm{dB}}(\omega_k). \tag{6.4}$$

The proposed frequency-dependent weighting factor, denoted $\alpha(\omega_k)$, is defined as

$$\alpha(\omega_k) = 1 - \left(\frac{\hat{\sigma}_{n,\mathrm{dB}}(\omega_k) - \min_k \hat{\sigma}_{n,\mathrm{dB}}(\omega_k)}{\Delta\hat{\sigma}_{n,\mathrm{dB}}} \right), \tag{6.5}$$

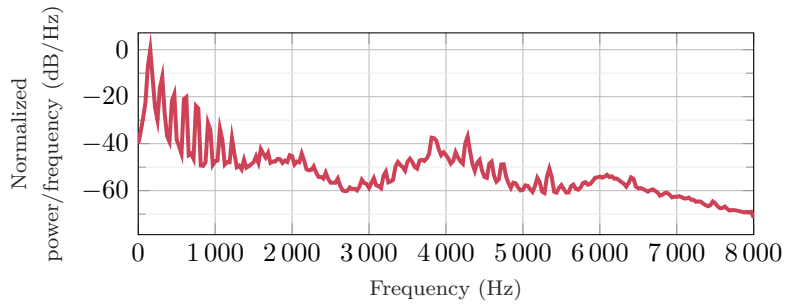
which assumes values between 0 and 1, as a function of the normalized PSD (in dB scale). The corresponding frequency-weighted broadband SRP, denoted $\tilde{z}(\gamma, \rho)$ is thus given by

$$\tilde{z}(\gamma, \rho) = 2 \sum_{p=1}^{P} \sum_{k=1}^{K} \Re\{\alpha(\omega_k)\psi_p(\omega_k) e^{j\omega_k \Delta t_p(\gamma,\rho)}\}. \tag{6.6}$$

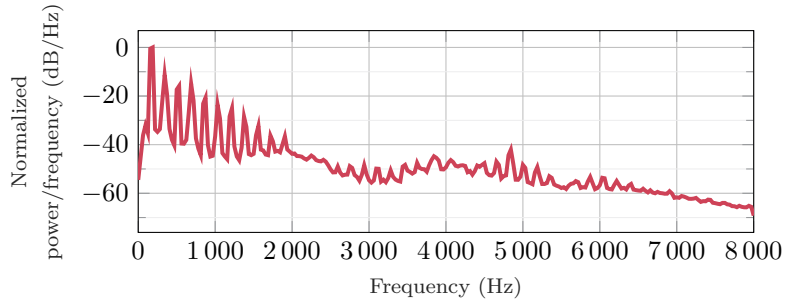
As an illustrative example, in Fig. 6.1, we present the normalized noise PSD, estimated from one of the propeller microphone signals[1], while the UAV was operating at constant thrust and simulated hovering condition, for averagee rotor speeds of 4000 rpm, 5000 rpm, and 6000 rpm. It can be observed that the ego-noise presents a visible harmonic structure, related to the blade passing frequency (BPF), and correspondingly, to the rotor speed [207]. In Fig. 6.2, we present the corresponding values of $\alpha(\omega_k)$, computed based on the estimated noise PSD for an average rotor speed of 5000 rpm, illustrated in Fig. 6.1b. The observed weighting structure indicates that the incorporation of $\alpha(\omega_k)$ in the formulation of $\tilde{z}(\gamma, \rho)$ is expected to penalize cross-correlation contributions at frequencies in which the ego-noise's harmonic components are expected to mask the target speech content.

The method proposed in **Ch. 3** can then be applied to estimate the speech source's DOA, while employing a weighted broadband SRP map, denoted $\tilde{\mathbf{z}}$ and computed based on (6.6). The angular grids used for constructing the SRP map $\tilde{\mathbf{z}}$ and the PSD vector $\boldsymbol{\varphi}_S$ are now designed for a set of azimuth and elevation pairs, rather than solely azimuth, as previously considered in **Ch. 3**.

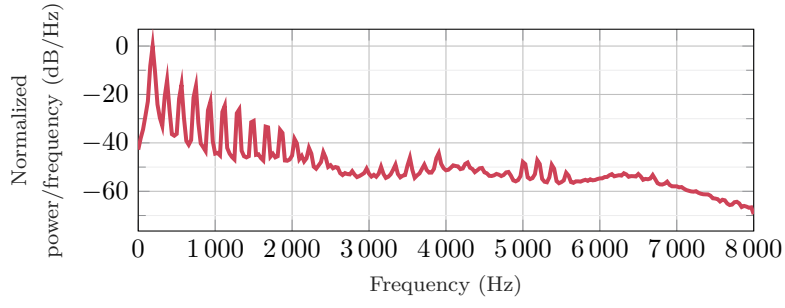
[1] Similarly to **Ch. 5**, the signal is downsampled from 44.1 kHz to 16 kHz. We estimate the signal's PSD via Welch's method, using a 512-point square-root Hann window corresponding to a length of 32 ms and 50% overlap. The signal's duration corresponds to 3 s, yielding a total of 187 frames.



(a) 4000 rpm



(b) 5000 rpm



(c) 6000 rpm

Figure 6.1: Estimated ego-noise PSD from recordings made with one of the propeller microphones, for different average rotor speeds.

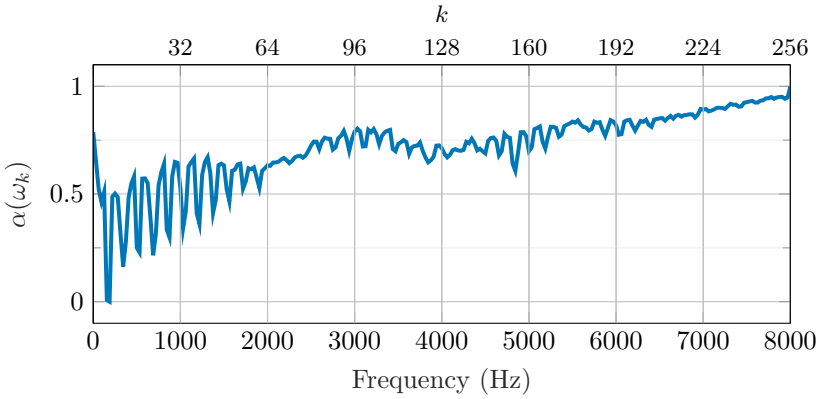


Figure 6.2: Weight values for average rotor speed of 5000 rpm.

6.3 Speech enhancement using ego-noise references and estimated source DOA

In the following equations, similarly to **Ch. 5**, we consider processing different frames and frequency bins independently, with their corresponding indices omitted for conciseness. We revisit the PK-MWF realization presented in Sec. 5.3.3, and simplify its expression as

$$\hat{\mathbf{w}}_{\text{PK-MWF}} = \hat{\mathbf{C}}\hat{\mathbf{w}}_{\text{GEVD}}^{\text{red}}\mathbf{H}^{\text{H}}\mathbf{e}_d, \tag{6.7}$$

with

$$\hat{\mathbf{w}}_{\text{GEVD}}^{\text{red}} = (\hat{\mathbf{Q}}^{\text{red}})^{-\text{H}}\text{diag}\left\{1 - \frac{\hat{\sigma}_{n1}^{\text{red}}}{\hat{\sigma}_{y1}^{\text{red}}}, 0, \ldots, 0\right\}(\hat{\mathbf{Q}}^{\text{red}})^{\text{H}}, \tag{6.8}$$

which corresponds to the GEVD-based implementation of an MWF filter [65, 83], applied to the microphone signals after the GSC stage represented by $\hat{\mathbf{C}}$. Hence, the MWF filter $\hat{\mathbf{w}}_{\text{GEVD}}^{\text{red}}$ is computed based on the reduced noise and speech-plus-noise correlation matrices' estimates, $\hat{\mathbf{R}}_{\mathbf{yy}}^{\text{red}}$, and $\hat{\mathbf{R}}_{\mathbf{nn}}^{\text{red}}$, respectively [83].

In the signal model presented in Sec. 5.2, one may note that the structure of the speech-only correlation matrix $\mathbf{R}_{\mathbf{ss}}$, assumed to be rank-1, is defined by the transfer function $\mathbf{a}$ from the speech source to the microphone array. Moreover, let $\mathbf{a}(\gamma, \rho)$ denote the relative transfer function (RTF) from a source in the far field, at a given azimuth γ and elevation ρ, to an M-element microphone array. If anechoic conditions are assumed, $\mathbf{a}(\gamma, \rho)$ corresponds to a direct-path RTF, and when normalized with respect to the array's first microphone, it is given by [6]

$$\mathbf{a}(\gamma, \rho) = [1,\, e^{j\omega_k \Delta t_{2,1}(\gamma,\rho)},\, \ldots,\, e^{j\omega_k \Delta t_{M,1}(\gamma,\rho)}]^{\top}. \tag{6.9}$$

Let $(\hat{\gamma}_s, \hat{\rho}_s)$ denote the target speech source's estimated azimuth and elevation pair. By assuming that the microphone array geometry is known, we propose to substitute the GEVD-based implementation of the MWF stage ($\hat{\mathbf{w}}^{\text{red}}_{\text{GEVD}}$) in the PK-MWF implementation in (6.7) by the alternative formulation of the MWF filter, comprising an MVDR beamformer followed by a single-channel Wiener filter [8, 63], in which the direct-path RTF, obtained based on the estimated DOA $(\hat{\gamma}_s, \hat{\rho}_s)$, can be incorporated. Hence, the PK-MWF is then reformulated as

$$\hat{\mathbf{w}}_{\text{DOA-PK-MWF}} = \hat{\mathbf{C}}\hat{\mathbf{w}}^{\text{red}}_{\text{DOA}}\mathbf{H}^{\text{H}}\mathbf{e}_d, \tag{6.10}$$

where

$$\begin{aligned}\hat{\mathbf{w}}^{\text{red}}_{\text{DOA}} = &\frac{\left(\hat{\mathbf{R}}^{\text{red}}_{\mathbf{nn}}\right)^{-1}\mathbf{a}^{\text{red}}(\hat{\gamma}_s, \hat{\rho}_s)}{\left(\mathbf{a}^{\text{red}}(\hat{\gamma}_s, \hat{\rho}_s)\right)^{\text{H}}\left(\hat{\mathbf{R}}^{\text{red}}_{\mathbf{nn}}\right)^{-1}\mathbf{a}^{\text{red}}(\hat{\gamma}_s, \hat{\rho}_s)} \\ &\times \frac{\hat{\phi}_s}{\hat{\phi}_s + \hat{\sigma}_n\left(\left(\mathbf{a}^{\text{red}}(\hat{\gamma}_s, \hat{\rho}_s)\right)^{\text{H}}\left(\hat{\mathbf{R}}^{\text{red}}_{\mathbf{nn}}\right)^{-1}\mathbf{a}^{\text{red}}(\hat{\gamma}_s, \hat{\rho}_s)\right)^{-1}},\end{aligned} \tag{6.11}$$

and

$$\mathbf{a}^{\text{red}}(\hat{\gamma}_s, \hat{\rho}_s) = \mathbf{H}^{\text{H}}\mathbf{a}(\hat{\gamma}_s, \hat{\rho}_s). \tag{6.12}$$

Note that the terms $\hat{\phi}_s$ and $\hat{\sigma}_n$ in (6.11) denote the data-dependent, estimated PSDs of the target source and noise, respectively. Moreover, in the preceding GSC stage (cf. (5.19)), an estimate of the noise-only correlation matrix $\hat{\mathbf{R}}_{\mathbf{nn}}$ is still required. Therefore, in this chapter, the proposed speech enhancement framework using the modified formulation of the PK-MWF is still not entirely data-independent, even though prior spatial information in terms of DOA is being employed. Given this context, it will be assumed that noise-only recordings are available for estimating $\hat{\mathbf{R}}_{\mathbf{nn}}$, and that the estimation of $\hat{\phi}_s$ and $\hat{\sigma}_n$ can be performed, for noisy speech recordings, using the method of instantaneous PSDs proposed in [230].

6.4 Evaluation

In order to evaluate the performance of the modified PK-MWF, we utilize the same set of experimental measurements detailed in Sec. 5.4.1, involving a UAV in simulated hovering mode and with average rotor speeds of 4000 rpm, 5000 rpm and 6000 rpm. In Sec. 6.4.1, the processing of the microphone recordings is further detailed for the tasks of DOA estimation and speech enhancement, and in Sec. 6.4.2, the results obtained are analyzed and compared with those previously presented in Sec. 5.5.

6.4.1 Processing

Similarly to Sec. 5.4.2, the measurements were processed offline by, firstly, downsampling the recorded signals from 44.1 kHz to 16 kHz. The short-time Fourier transform (STFT) of the microphone signals is computed using a 512-point square-root Hann observation window with 50% overlap.

In the DOA estimation stage, it is assumed that the first two seconds of the recorded microphone signals only contain the ego-noise component, such that the weighting function $\alpha(\omega_k)$ can be computed based on a pre-estimation of the noise PSD from one of the propeller microphones, as detailed in Sec. 6.2.2. We employ $M = M_{S+N} = 12$ microphones available from the main array, yielding a total of $P = 66$ distinct microphone pairs, to perform the DOA estimation over three seconds of noisy speech recordings, corresponding to a total of 187 STFT frames. The cross-correlations used in the SRP function in (6.6) are obtained by averaging the STFT cross-correlations, and a total of $K = 256$ frequency bins are used.

The angular grids used to build the weighted broadband SRP map $\tilde{\mathbf{z}}$ and the PSD vector $\hat{\boldsymbol{\varphi}}_S$, respectively, are defined over a spherical cap, i.e., a region of a sphere, for elevation angles between $-30°$ and $-90°$. The azimuth is considered over the full range of $360°$ [2]. Due to the significant increase in computational complexity, related to the inclusion of the target source's elevation into the DOA estimation problem, the resolution of both grids is limited to $15°$ in azimuth and elevation, such that a total of $N = L = 97$ candidate DOAs are considered.

Once the estimate of the source's DOA pair $(\hat{\gamma}_s, \hat{\rho}_s)$ is obtained by employing the DOA estimated method from **Ch. 3**, while employing the weighted broadband SRP map $\tilde{\mathbf{z}}$, the corresponding direct-path RTF $\mathbf{a}^{\text{red}}(\hat{\gamma}_s, \hat{\rho}_s)$ is computed based on (6.9) and (6.12), and incorporated into $\hat{\mathbf{w}}^{\text{red}}_{\text{DOA}}$ as in (6.11). The noise correlation matrices $\hat{\mathbf{R}}_{\mathbf{nn}}$ and $\hat{\mathbf{R}}^{\text{red}}_{\mathbf{nn}}$, employed in (5.20) and (6.11), respectively, are estimated based on the first two seconds of noise-only recordings. In the noisy speech recordings, the target source and noise PSDs $\hat{\phi}_s$ and $\hat{\sigma}_n$, respectively, are estimated for each frame by using the method of instantaneous PSDs for speech enhancement from [230].

Differently from **Ch. 5**, we limit the evaluation of the speech enhancement framework performance to the case where all $M_{S+N} = 12$ microphones from the drone's main array are used in the PK-MWF implementation, with the $M_N = 4$ propeller microphone signals used as ego-noise references. The results obtained, when considering the three different rotor speeds, are compared with those previously presented in Sec. 5.5, concerning the PK-MWF implementation

[2] Due to the $-90°$ elevation corresponding to the bottom of the spherical cap defining the angular grid, a single candidate direction is defined for this elevation angle.

Table 6.1: Absolute DOA estimation errors for different average rotor speeds.

Average rotor speed	Absolute error (Azimuth)	Absolute error (Elevation)
4000 rpm	42°	5°
5000 rpm	8°	5°
6000 rpm	8°	5°

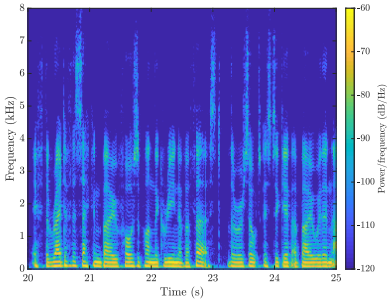
using the SPP estimated from the embedded microphone (iSPP) and from the reference external microphone (xSPP). The comparison is performed through an analysis of the resulting spectrograms obtained with each framework, as well as of the SNR, STOI and PESQ improvements.

6.4.2 Results

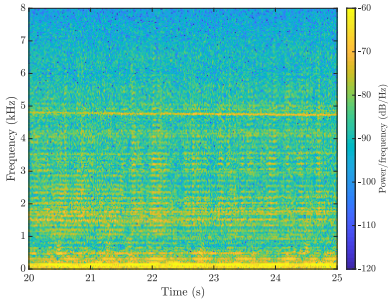
The absolute errors obtained in the DOA estimation stage, for different average rotor speeds, are presented in Table 6.1. It can be observed that the azimuth estimation was particularly less robust when the average rotor speed was 4000 rpm. This result could be due to the presence of a greater number of distinct harmonic components in the ego-noise, related to the BPF being lower than in the cases of 5000 rpm and 6000 rpm. However, further investigation would be required to consistenly evaluate the DOA estimation performance in the context of UAV recordings.

In Fig. 6.3, the spectrograms of a 5-second segment of the clean reference speech signal, as well as of the speech signal corrupted by ego-noise and the resulting outputs of each approach considered in the comparison are presented. The average rotor speed corresponds to 5000 rpm. It can be observed that using the proposed DOA-based modification of the PK-MWF filter contributes to removing more noise than the PK-MWF implementation using iSPP, for frequencies above approximately 1500 Hz. However, for frequency bands below that value, the result obtained with the DOA-based PK-MWF is seemingly still more corrupted by residual ego-noise than when the PK-MWF with iSPP is used. The result obtained with the PK-MWF implementation using xSPP, when compared to those obtained using iSPP and the DOA-based approach, again indicates the importance of properly estimating the correlation matrices in such MWF-based approaches, as previously noted in Sec. 5.5.

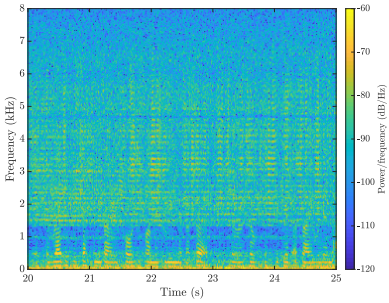
After conducting a visual analysis of the spectrograms depicted in Fig. 6.3 and



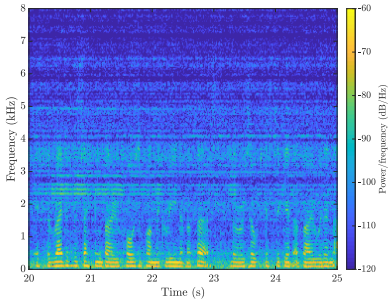
(a) Clean speech signal



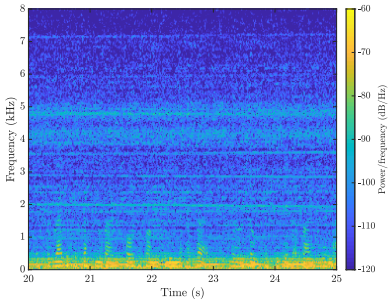
(b) Noisy speech signal



(c) PK-MWF with iSPP

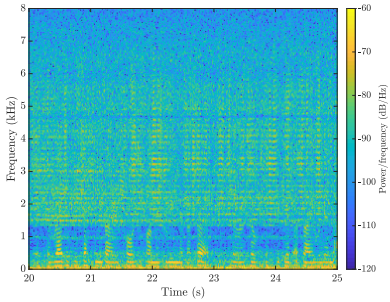


(d) PK-MWF with xSPP

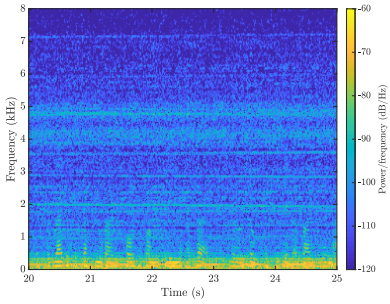


(e) PK-MWF with DOA

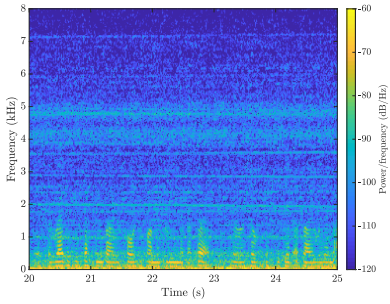
Figure 6.3: Spectrograms of signals recorded with the reference channel of the UAV's microphone array, as well as the corresponding outputs of the compared methods, for an average rotor speed of 5000 rpm.



(a) PK-MWF with iSPP



(b) PK-MWF with DOA



(c) PK-MWF with iSPP and DOA

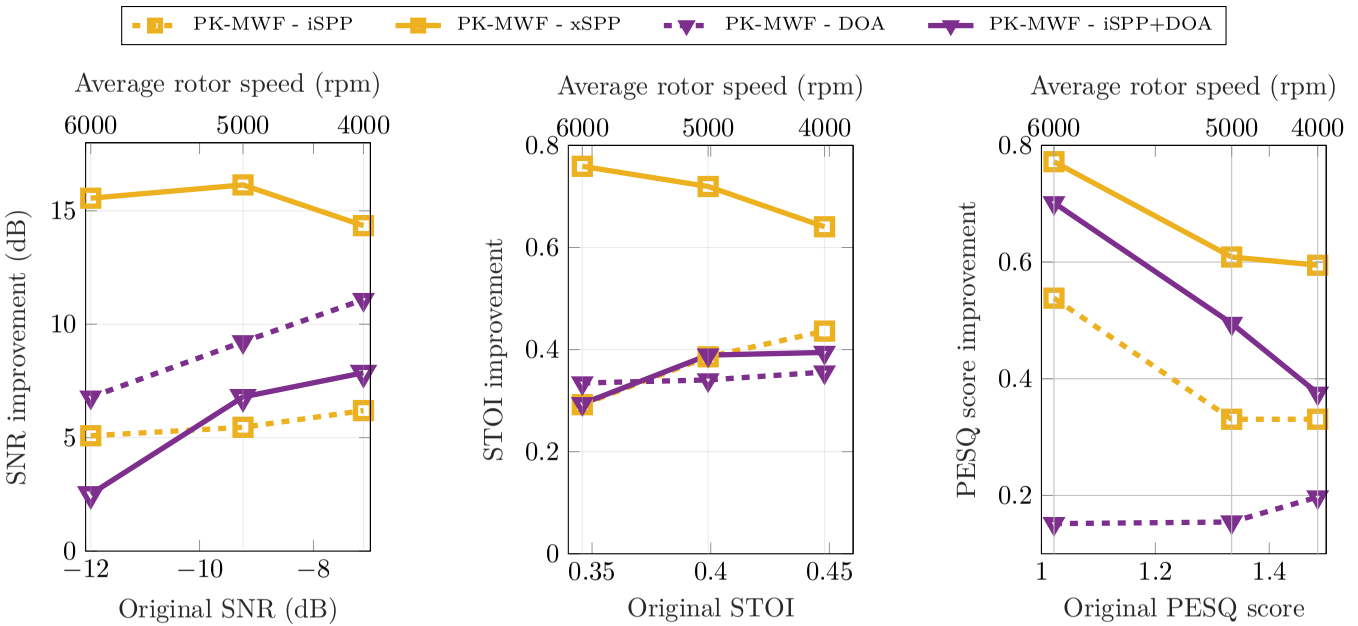
Figure 6.4: Spectrograms of output signals when employing the different implementations of PK-MWF being compared, for an average rotor speed of 5000 rpm.

considering the observations on the spectral characteristics of ego-noise detailed in Sec. 6.2.2, we investigate an alternative speech enhancement approach. We consider using the iSPP to estimate the correlation matrices and employ the PK-MWF formulation in (6.7) for frequencies below 1500 Hz, while the newly modified formulation in (6.10), based on the estimated DOA, is employed for frequencies above 1500 Hz. In Fig. 6.4, the resulting spectrogram from the combined application of both methods is presented, in comparison to the results obtained with each method applied across all frequency bins. It can be observed that this proposed alternative approach could potentially yield an improvement in SNR in comparison to the PK-MWF using iSPP across all frequency bins. Additionally, the lower-frequency components of speech are potentially better preserved when compared to the sole application of the DOA-based implementation across all frequency bins.

In order to better assess the performance improvement, and similarly to what is presented in Sec. 5.5, we compute the SNR, STOI and PESQ improvements obtained by employing the DOA-based implementation of the PK-MWF across all frequencies, as well as the alternative approach combining iSPP and DOA estimates. The results are compared to those obtained when employing the PK-MWF solely based on iSPP or xSPP. In Fig. 6.5, it can be observed that the DOA-based implementation of PK-MWF provides a significant SNR improvement in comparison to the PK-MWF using iSPP, albeit at the expense of the perceptual improvement in terms of STOI and PESQ. Interestingly, it can also be observed that combining the DOA and iSPP-based PK-MWF implementations can improve performance in terms of PESQ, while the STOI improvement is mostly similar. In that case, however, the SNR improvement is lower than the one obtained with the implementation solely based on the estimated DOA.

6.5 Conclusion

In this chapter, the practical application of estimated DOA information was investigated within the context of speech enhancement with a microphone array embedded in a UAV. This was achieved by modifying the PK-MWF filter implementation, initially presented in **Ch. 5**, which relied solely on SPP estimates. In the proposed modification, an estimated source DOA, in terms of azimuth and elevation, is obtained through the methodology presented in **Ch. 3**, while employing a weighted broadband SRP map that is computed while considering the specific spectral characteristics of the UAV ego-noise. Then, the corresponding direct-path RTF could be computed and integrated into the framework's MWF stage, which is reformulated as an MVDR beamformer followed by a single-channel Wiener filter. Results obtained from experimental recordings demonstrated that employing the DOA-based implementation of the PK-MWF filter significantly improves the SNR of the filtered signal when compared to the result obtained when employing the PK-MWF with SPP estimated from one of the embedded microphones, whereas the perceptual improvement, in terms of STOI and PESQ, is compromised. Additionally, it was observed that combining both DOA and SPP-based implementations of PK-MWF can improve performance in terms of PESQ, while the SNR improvement is lower than the one obtained when only employing estimated DOA information.



(a) SNR improvement

(b) STOI improvement

(c) PESQ improvement

Figure 6.5: Performance improvement in terms of different objective measures associated with different average rotor speeds ($M_{\mathrm{S+N}} = 12$).

Chapter 7

Research valorisation plan

In this chapter, a strategic analysis of the academic contributions in this thesis is presented by means of a research valorisation plan. This plan aims to outline how the research outcomes presented in this thesis can be transferred into an industrial context. Additionally, it aims to highlight the potential socio-economic value that could be generated through such process.

This chapter is organized as follows. In Sec. 7.1, the main outcomes of the scientific research presented in this thesis are categorized. In Sec. 7.2, the collaborations with industry partners that took place during the PhD and contributed to guiding the contributions presented in the thesis are outlined. In Sec. 7.3, a brief analysis of the current audio market is presented. In Sec. 7.4, the validation aspects for the contributions in this thesis, while considering practical applications, are discussed.

7.1 Research outcomes

This section categorizes the research findings discussed in the preceding chapters of this thesis into methods and dataset. Note that the published research outcomes are available in Open Access under the corresponding copyright licensing agreements specified below, contributing to the dissemination of the presented academic research and its potential impact on future developments and collaborations.

7.1.1 Methods

The research outcomes in terms of methods are as follows.

In **Ch. 2**, a method for direction-of-arrival (DOA) and power spectral density (PSD) estimation using a single, directional and rotating microphone is presented, applicable in scenarios involving single or multiple sound sources. This chapter's related paper has been published as an open-access, international peer-reviewed journal paper [124], under a CC BY license.

In **Ch. 3,** a method for direction-of-arrival (DOA) estimation using a multi-microphone setup is presented. This chapter's related paper has been submitted for publication.

In **Ch. 4**, a method for an experimental measurement protocol for evaluating UAV noise emissions is presented. This chapter's related paper has been published as an open-access, international peer-reviewed paper [126], under a CC BY license.

In **Ch. 5**, a method for speech enhancement with a microphone array embedded in an unmanned aerial vehicle (UAV) is presented. This chapter's related paper has been published in an open-access, international peer-reviewed conference proceedings [127], under a CC BY-NC license.

In **Ch. 6**, an extension of the speech enhancement method from **Ch. 5** is presented. This contribution has not been submitted elsewhere.

7.1.2 Dataset

In **Ch. 5**, along with a contribution in terms of methods for speech enhancement using ego-noise references with a microphone array embedded in a UAV, an open-access dataset containing the recordings made with the embedded microphone array is made available online [123], under a CC BY-NC-SA license.

7.2 Research collaborations with industry partners

The direction of the academic research presented in this thesis has been greatly influenced by the collaborations with different industry partners throughout the PhD. In this section, we outline these collaborations in terms of their application context, main objectives and outcomes.

7.2.1 VLAIO R&D Project no. HBC.2017.0358 "SPOTT - Tomorrow's Scalable and PersOnalised advertising Technology, Today".

Duration of participation: April 2019 to February 2020.

Industry partner: Appiness NV

Research partners:

- UGent - imec - IDLab (Image & Video Processing and Speech & Audio Processing, Faculty of Engineering & Architecture - ELIS)
- **KU Leuven - Department of Electrical Engineering (ESAT) - Center for Dynamical Systems, Signal Processing and Data Analytics (STADIUS)**
- KU Leuven - Department of Electrical Engineering (ESAT) - Processing Speech and Images (PSI)

Application context and objectives: The SPOTT mobile application, from Appiness NV, works as a video content enrichment platform. It allows for the identification of characters and products featured in a wide range of TV programs, while enabling the mobile device being used to serve as a second screen. The usage of the SPOTT app serves the dual-purpose of

a) facilitating the online purchase of products showcased on the TV program being watched, and

b) retrieving additional information about the identified products.

In this interdisciplinary R&D project, the industrial partner Appiness NV aimed to improve the backend architecture and workflow of the SPOTT app. In order to properly identify the programs being shown on TV, the SPOTT

app initially relied on third-party audio fingerprinting technology to match the fingerprint of an audio excerpt from the content being played with entries from the database of available audio fingerprints. Although such an approach has shown to be effective in music retrieval applications, the scenarios in which SPOTT is assumed to be used present additional challenges. These scenarios involve predominantly speech signals, such as character dialogues, TV presenters, and more, and may be subject to interference from various noise sources typically present in a home environment, such as additional speakers, background chatter, and household appliances. Hence, in this project, the role of the STADIUS research team was to perform a noise and reverberation characterization of recordings made in a home environment, such that, rather than focusing on denoising or dereverberation techniques to be applied over the recorded signals to maintain the performance of the conventional audio fingerprinting framework being used, these acoustic features could be included in a newly developed audio template matching algorithm that compares the observed microphone signals' features with the database features.

Collaboration outcomes: Throughout the project's duration, valuable experiences have been acquired with respect to the deployment of audio analysis techniques in practical scenarios. Unlike the conventional use of microphone array processing techniques for spatial audio analysis, the setup limitations encountered during the project led to the decision of focusing on a single-microphone spatial analysis approach, with the purpose of performing noise and reverberation characterization. By assuming that the single microphone available is directional, allowed to rotate, and its orientation can be inferred from an additional sensor (e.g., a smartphone built-in gyroscope), a first method for DOA and PSD estimation was proposed, and at a later stage, extended to the contribution presented in Ch. 2. The participation in this project provided valuable insights into the potential applications of single-microphone audio analysis techniques. Furthermore, it provided the opportunity to gain an insider's perspective on the specific requirements and limitations of practical audio-based solutions, indicating the current challenges in this application context.

7.2.2 Experimental measurements with an unmanned aerial vehicle (UAV)

Duration of participation: May 2021 to July 2021.

Industrial partner: Siemens Digital Industries Software

Research partners:

- KU Leuven - Department of Mechanical Engineering - Mecha(tro)nic System Dynamics (LMSD)
- **KU Leuven - Department of Electrical Engineering (ESAT) - Center for Dynamical Systems, Signal Processing and Data Analytics (STADIUS)**

Application context and objectives: With a growing interest in expanding the range of applications in which UAVs can be used, recent efforts have been focused on evaluating the noise emission characteristics of these systems, as well as on investigating the opportunities in UAV-specific audio processing applications, such as auralization of fly-over noise, and on-board audio recording. To this end, this collaboration aimed at designing an experimental measurement protocol for evaluating UAV noise emissions, as well as developing a speech enhancement method involving a microphone array embedded in a UAV.

Collaboration outcomes: Experimental measurements were organized and executed with a customized UAV, pertaining to STADIUS, in a semi-anechoic chamber at Siemens Digital Industries Software. Audio recordings were performed with external microphones, arranged in a hemispherical array configuration, as well as with a microphone array embedded in the customized UAV. These recordings were captured while the UAV was mounted on a stand, simulating a hovering state with active rotors, and used to conduct a UAV ego-noise emission study. Furthermore, the measurements were used to experimentally validate a speech enhancement framework, with its development influenced by the UAV's embedded array configuration. From the active participation in designing and conducting the measurement campaign to employing the audio recordings acquired, this research collaboration contributed to broadening the understanding of UAV ego-noise characteristics, as well as the opportunities and challenges in on-board audio recording and processing.

The involvement of collaborators from different technical backgrounds provided relevant insights on the work conducted from both an acoustics-based and an audio signal processing-based perspective. Hence, this collaboration has contributed to a comprehensive understanding of the application context,

by bridging these distinct yet complementary fields. Furthermore, the collaboration's industry-oriented nature provided a practical opportunity to further understand the priorities concerning methodology, setups, and other technical aspects pertinent to this particular domain.

7.3 Strategic analysis of relevant markets

In order to illustrate the potential socio-economic value that the contributions presented in this thesis can generate, an analysis of the current market encompassing different applications is presented. This analysis focuses on the main sectors where the contributions in this thesis are considered relevant. Furthermore, it emphasizes the economic growth trends within the main sectors, as well as the influential factors driving their expansion.

- **Hearing aids:** The World Health Organization (WHO) has reported that hearing loss is currently affecting over 1.5 billion individuals globally [231]. When left untreated, moderate to severe hearing loss has been linked to negative impacts on communication skills, mental health, as well as employment and educational opportunities [231]. Consequently, the demand for hearing aids continues to rise, despite existing technologies not being readily affordable in numerous countries [232]. As of 2022, the hearing aid market was estimated at a value of 7.3 billion USD, demonstrating a compound annual growth rate (CAGR) of 5.6 % from 2023 to 2032 and a projected value of 12.6 billion USD within the forecast period [233]. Within this context, the source localization and speech enhancement methods presented in this thesis could contribute to improving hearing aid technology, and therefore, the listening comfort and overall quality of life of its users. For instance, the single-microphone DOA and PSD estimation method could be employed while exploiting natural head movements of a hearing aid user.

- **Automotive audio:** In recent years, the demand for enhanced hearing experience, as well as the evolution of autonomous vehicles, have influenced the automotive audio market [234]. Speech enhancement and source localization methods can be used for in-cabin communication and voice command systems [235–237]. Additionally, in autonomous cars, although navigation is mostly based on information from camera, laser and radar sensors [107], audio analysis techniques can still be exploited in a complementary redundancy system, which is of great importance in such a safety-critical context [107, 109]. The automotive audio market was valued at 10 billion USD in 2022, with a CAGR of 5 % for the forecast

period from 2023 to 2032, with a projected value of 20 billion USD in 2032 [234].

- **Spatial audio:** In applications such as virtual reality (VR), augmented reality (AR), and soundscape design, spatial audio analysis and reproduction play a crucial role in enhancing their immersive aspect [238, 239]. Despite many advancements in visual augmentation systems, challenges in spatial audio technology for immersive applications still remain present [238, 240], indicating development opportunities within this domain. The spatial audio market was valued at 5 billion USD in 2022, with a CAGR of 10 % between 2023 and 2032, with a value projected at 15 billion USD in 2032 [241].

- **Automatic speech recognition:** The use of automatic speech recognition (ASR) techniques has shown to improve human-machine interfaces by enabling functionalities such as digital assistance and voice-based search [92], adding tangible social value in terms of communication and accessibility [242, 243]. Some relevant applications of ASR include smartphones, smart homes and in-vehicle communication [9, 93–98]. Furthermore, the use of source localization and speech enhancement techniques has shown to improve the performance of ASR systems [9, 244]. The ASR market was valued at USD 10.42 billion in 2022, with an anticipated CAGR of 24.8% from 2023 to 2030 [245].

- **Smart home systems:** In smart home systems, audio sensors have become the most natural way for residents to interact with the smart home system through ASR [100], while simultaneously allowing for the detection and classification of events [11]. Therefore, performing these tasks can benefit from speech enhancement and source localization methods. The smart home market was valued at 80.2 billion USD in 2022, with a CAGR of 20.1 % from 2023 to 2030, with the forecasted value of 338.28 billion USD in 2030 [246].

- **Videoconferencing:** The demand for improved audiovisual quality and functionality in videoconferencing systems aligns with the evolving tendencies of remote work and virtual collaboration [247, 248]. As such, the videoconferencing market is expected to witness further innovation and development in technologies aimed at improving the overall communication experience. Examples of multi-party conferencing functionalities that can benefit from the source localization and speech enhancement contributions presented in this thesis include camera-steering, speaker separation and noise reduction. The videoconferencing market was valued at 25 billion USD in 2022, with a CAGR at 10 % between 2022 and 2032, and its projected value in 2032 is of 95 billion USD [248].

- **Drone services:** Continuous advancements in UAV technology have contributed to their use becoming more accessible and versatile. The range of applications in which they are employed has broadened from the initial military realm, to other sectors such as filmmaking, entertainment, agriculture and delivery services [249]. Despite most applications still relying on visual data [250], the field of 'drone audition' has recently gained traction [110], due to its potential to be employed as a complementary modality when image or video capturing is unreliable [110, 250]. Different techniques have been investigated to perform sound source localization and speech enhancement under the adverse acoustic conditions that drone applications normally impose [110, 111]. The drone services market was valued at 10.88 billion USD in 2020, with a CAGR at 24 % between 2021 and 2028, reaching a projected value of 60.7 billion USD in 2028 [251].

7.4 Validation

In order for the contributions presented in this thesis to be properly exploited and validated in an industrial setting, it is essential to consider different aspects to ensure the reliability and applicability of the research findings. In this section, we present the validation aspects considered and the corresponding follow-up actions needed with respect to the different contributions presented in this thesis.

Robustness: In order to ensure reliable performance of any audio processing method in practical scenarios, an extensive validation procedure is required. In Part I, although the performance of the single and multi-microphone source localization methods presented has been extensively evaluated through a series of simulations, it is crucial to employ experimental data in the validation process within an industrial context. By doing that, the nuances of the practical scenarios considered for a given targeted application can be analyzed, such that the methodologies' robustness is properly evaluated. In the case of the single-microphone DOA and PSD estimation methods from Ch. 2, these nuances are believed to be mainly related to challenges such as non-stationarity of both target and noise components, as well as nonlinearities in the single microphone response. Additionally, it would be required to evaluate the methods' robustness when considering uncontrolled microphone rotations, which are more likely to occur in practical applications such as hearing aids and smartphones. Regarding the multi-microphone DOA estimation method from Ch. 3, influencing factors pertaining to practical scenarios such as the microphone array geometry, calibration and synchronization of signals, are believed to require further investigation in the validation process. In Part II, although the proposed

experimental setup from Ch. 4 has provided insights into UAV ego-noise emission characteristics, further steps are believed to be required in order to effectively investigate the spatial symmetry of the UAV ego-noise emissions (cf. [126]). This could involve reassessing the distribution of microphones, and possibly extending the configuration to a fully spherical microphone array. Moreover, although experimental measurements have been performed with a microphone array embedded in a fixed-position UAV and used in the evaluation of the multi-microphone speech enhancement methods from Ch. 5 and Ch. 6, it is also essential to consider scenarios with a flying UAV. Understanding the effect of spatially dynamic recordings and accounting for factors like wind noise are believed to be the main steps towards the further development of these methods within this application context.

Implementation: Throughout the doctoral research, the methods presented in this thesis were implemented and tested using MATLAB, a higher level programming language commonly used in the stages of prototyping and algorithmic development of digital signal processing solutions. While MATLAB provides a convenient framework in these initial stages, porting the algorithms proposed in this thesis to lower-level languages, such as C, C++, or assembly language, becomes imperative for industrial applications. Moreover, a seamless integration of the newly developed digital signal processing modules into the software architectures of interest is required.

Scalability and real-time processing: When implementing and applying the different methods proposed in this thesis, it is important to assess the resulting framework's scalability within a specific application context. In Ch. 2 and Ch. 3, a brief computational complexity analysis is presented for the single and multi-microphone source localization methods proposed, respectively. However, further validation is still required to evaluate their efficiency in practical setups, allowing for informed decisions to be made when considering the framework's design and resource allocation.

Furthermore, the algorithm development in all contributions presented in this thesis involved offline processing of microphone signals. However, considering the potential applications in which the different source localization and speech enhancement methods proposed could prove valuable, it is also necessary to explore achieving real-time processing implementation alternatives. For instance, in environments characterized by dynamic acoustic conditions or moving microphone setups, it is expected that the DOA estimation methods presented in Ch. 2 and Ch. 3 could be efficiently combined with tracking methods in order to avoid successively running the iterative algorithms that are proposed. Within the context of UAV applications, as considered in Ch. 5 and Ch. 6, particularly time-sensitive tasks, such as search and rescue or autonomous navigation, further evaluation of real-time processing capabilities is required.

Chapter 8

Conclusion

This thesis explored challenges within the domain of source localization and speech enhancement by presenting and discussing a collection of research contributions. The contributions primarily focused on addressing challenges emerging from limitations encountered in different applications, related to constrained microphone setups and adverse acoustic conditions.

The remainder of the conclusion is organized as follows. In Sec. 8.1, the contributions presented in this thesis are summarized. In Sec. 8.2, a few suggestions for future research are discussed.

8.1 Summary

In **Part I**, the contributions on the subject of source localization were presented. Specifically, in **Ch. 2**, the problem of single-channel source localization was addressed, motivated by the limitations in spatial audio analysis capabilities imposed by single-microphone setups, or devices in which simultaneous access to multiple microphone signals is unreliable. Two approaches were presented for estimating the direction of arrival (DOA) and power spectral density (PSD) of stationary point sources by using a single, rotating, directional microphone. Both approaches were based on the premise that a directional microphone will capture a spatially static and localized sound source with a different response depending on its orientation. Hence, as the microphone orientation varies for different observation time frames, changes in the microphone signal PSD can be analyzed for estimating the DOAs of the sources and their corresponding PSDs.

In both proposed approaches, the solution of a group-sparsity-regularized optimization problem allowed for estimating direction-dependent PSD values relative to a given angular dictionary, and locating peaks in the estimated PSD vector, which in turn provided the DOA estimates. The first approach, referred to as the Group Lasso Least Squares (GL-LS) approach, was based on the use of a broadband signal model for formulating the group-sparse optimization problem, and the DOA estimation was followed by the re-estimation of the source PSDs through the solution of an overdetermined least squares problem. The second proposed approach, referred to as the Group Lasso with l_1-reweighting (GL-L1), alternatively employed a narrowband signal model when building the linear system of equations used for formulating the group-sparse optimization problem, where the DOAs and PSDs could be jointly estimated by iterative, group-wise reweighting. In both GL-LS and GL-L1, the number of observations made with the rotating directional microphone can be lower than the total number of candidate DOAs in GL-L1.

The single-channel DOA and PSD estimation approaches were implemented using the alternating direction method of multipliers (ADMM), and their performance was evaluated and compared through an extensive set of simulations. Based on the results obtained, it was observed that the GL-L1 approach can present an advantage over GL-LS in terms of DOA estimation performance in scenarios with low SNR or where multiple sources are closely located to each other, indicating the benefit of the narrowband model and reweighting scheme in these cases. Moreover, it was shown that having the least squares PSD re-estimation step is beneficial in most scenarios tested, such that GL-LS outperformed GL-L1 in terms of PSD estimation errors.

In **Ch. 3**, we considered the problem of multi-channel source localization. Although the setup was not assumed to be limited to a single microphone in the newly considered context, the methodology from the first contribution presented in **Ch. 2** was exploited, motivated by the insights acquired regarding the possibility to locate multiple sources with a limited number of microphone orientations. Thus, in **Ch. 3**, a method was proposed for estimating the direction of arrival (DOA) of multiple broadband sound sources. This was achieved through the solution of a group-sparse optimization problem, which modeled an observed broadband steered response power (SRP) map as a linear combination of power spectral densities (PSDs). Similarly to **Ch. 2**, the estimation of the source DOAs was then accomplished by identifying the peaks within the estimated PSDs integrated over frequency. An implementation using the alternating direction method of multipliers (ADMM) was presented, and simulation results show that when using SRP maps with lower grid resolutions with respect to the sampling requirements of the broadband SRP function, the proposed method outperformed the conventional SRP-PHAT method in the

multi-source scenarios tested. It was also shown that the proposed method performed similarly to the group-sparse iterative covariance-based estimation (GSPICE) method, here applied to estimate the source DOAs by directly fitting the microphone array cross-correlations into the candidate source PSD model used, at a lower computational cost. Finally, it was observed that the proposed method outperformed both SRP-PHAT and GSPICE in the cases of closely located sources.

In **Part II**, the contributions on the subject of speech enhancement were presented. The specific scenario of speech enhancement with a microphone array embedded in a UAV was considered, which entailed highly adverse acoustic conditions, due to the strong presence of ego-noise. Based on the source localization contributions presented in **Part I**, we focused on investigating the feasibility of employing spatial filtering-based methods for speech enhancement in such a challenging scenario.

Firstly, motivated by the need to further understand the spatial and spectral characteristics of UAV ego-noise, an experimental methodology was proposed for measuring ego-noise emissions in **Ch. 4**. A hemispherical microphone array setup was used, which allowed for simultaneous noise capture across various elevation and azimuth angles, and the methodology was demonstrated using a quadrotor drone, operating at constant thrust and simulated hovering condition. An analysis on the noise emission levels was performed and the results obtained revealed a consistent rise in sound pressure levels (SPLs) across all observed elevation angles, with increasing rotor speed. Additionally, through the analysis of the SPLs in one-third octave bands, it was observed, for selected rotor speeds, that while the drone noise emission levels were similar for different elevation angles, some differences in spectral characteristics were present.

In **Ch. 5**, a first contribution addressing the challenge of speech enhancement with a microphone array embedded in a UAV was presented. Although this contribution did not rely on knowledge of the target speaker's location, it still exploited spatial information provided by the microphone array configuration being used. Specifically, ego-noise reference signals were captured with microphones located near the UAV's propellers and used in the prior knowledge multichannel Wiener filter (PK-MWF) [83]. Speech presence probability (SPP) was estimated for detecting speech activity from an external microphone near the speech source, providing a performance benchmark, as well as from one of the embedded microphones, assuming a more realistic scenario. The estimated SPP was then used in the estimation of the noise-only and speech-plus-noise correlation matrices, required in the GEVD-based implementation of the PK-MWF. Experimental measurements were performed in a semi-anechoic chamber, with a UAV mounted on a stand and a loudspeaker playing a speech signal, while setting three distinct and fixed propeller rotation speeds, resulting in three

different SNRs. The recordings obtained and made available online [123] were used to compare the proposed method to the use of the standard multichannel Wiener filter (MWF) estimated with and without the propeller microphones being used in its formulation. Results showed that the use of PK-MWF while employing the ego-noise references achieved higher levels of improvement in speech intelligibility and quality, measured by STOI and PESQ, while the SNR improvement was similar.

In **Ch. 6**, the use of DOA information within the context of speech enhancement with a microphone array embedded in a UAV was investigated. The PK-MWF filter implementation presented in **Ch. 5** was modified, such that the computation of the MWF stage did not rely on a speech activity estimate. Instead, an estimated target source DOA was used to obtain a direct-path RTF, which in turn was incorporated through the alternative formulation of an MVDR beamformer followed by a single-channel MWF post filter. In order to test the proposed modification in the speech enhancement framework, the target source's DOA was estimated, in terms of azimuth and elevation, through the methodology presented in **Ch. 3**, while processing the same experimental recordings as those used in **Ch. 5**. Results obtained demonstrated that employing the DOA-based implementation of the PK-MWF filter significantly improves the SNR of the filtered signal when compared to the result obtained when employing the PK-MWF with SPP estimated from one of the embedded microphones, whereas the perceptual improvement, in terms of STOI and PESQ, is compromised. Additionally, it was observed that combining both DOA and SPP-based implementations of PK-MWF can improve performance in terms of PESQ.

8.2 Future research

Given that the contributions presented in this thesis were motivated by practical aspects posing challenges in the tasks of sound source localization and speech enhancement, the research outcomes presents numerous potential applications in an industrial context. As outlined in **Ch. 7**, the methodologies targeting different research problems, as well as the data acquired and made available online, can be exploited in the development of tailored solutions for a wide range of audio markets showing increase in demand. Nevertheless, it is essential to consider different aspects in future work to ensure the reliability and applicability of the research findings. Some examples on possible future research are presented as follows.

- The single-channel source localization methods from **Ch. 2** were developed and tested while considering a uniform distribution of orientations for

acquiring the microphone observations, which, in practice, would require the device to be rotated in very specific directions. Since in many applications, such as hearing aids or smartphones, the device's rotation cannot be fully controlled, it is fundamental to investigate the current method's performance limitations in such realistic cases. In future research, a possible approach to particularly address these cases could be by promoting an asymetric directivity pattern for the microphone being used, which could be achieved, for instance, by using a scattering structure, such that the necessary spatial information can still be captured. Another possibility to address uncontrolled microphone movements would be to exploit and adapt data augmentation and array interpolation methods to the case of a single rotating directional microphone [156, 252–254].

- The multi-channel source localization from **Ch. 3** was developed with a fixed grid resolution for both the observed broadband SRP map and the spatial power density, constructed based on location-specific PSDs. A possibility to reduce computational complexity, as well as to improve performance, consists in developing a grid refinement approach, such that a coarse grid could be initially used to identify the regions of interest, and then be iteratively refined to estimate the DOA of multiple sources. Additionally, since it was observed that the proposed method's performance was particularly dependent on the level of reverberation, this acoustic condition could be accounted for in the formulation of the optimization problem. Moreover, the evaluation of the multi-channel source localization method from **Ch. 3** requires to be extended to different microphone configurations, and the effect of issues in array calibration and synchronization should also be considered when targeting practical applications.

- The experimental methodology for measuring ego-noise emissions of a UAV from **Ch. 4** could be adapted by reassessing the distribution of microphones in the array configuration used, such that the measurements could be properly evaluated for a wider range of azimuth and elevation angles. The array configuration could also possibly be extended to a fully spherical microphone array around the UAV. Regarding other operational modes of the UAV other than hovering, the experimental methodology would also require further adaptation, such that it allows for performing measurements for, e.g., take-off and forward flight conditions. Additionally, sensors of different modalities such as cameras, real-time location systems (RTLS), and if outdoors, GPS systems, could be employed in conjunction with the microphone setup to allow for further analysing the relation between the ego-noise emissions and flight conditions.

- The method proposed for performing speech enhancement with a microphone array embedded in a UAV, presented in **Ch. 5**, was implemented while the UAV was in simulated hovering position and the ego-noise was assumed to be stationary during the entire observation. Hence, the correlation matrices employed in the formulation of the PK-MWF filter were estimated once and offline, based on all observation frames. Consequently, when considering more realistic flight conditions, in which the need for stability in the UAV or varying movements results in changes in rotor speed, it is necessary to extend the method's implementation, such that the estimates of the correlation matrices, and consequently, the filter coefficients, can be successively updated. Additionally, if external information on the rotor's speed is available, it could be incorporated as a parameter in the process of detecting when and how often the estimates need to be updated, as an attempt to partly circumvent the unavoidable increase in computational cost. When considering outdoor applications of the proposed framework, the effect of wind noise could be accounted for in the signal model, since it can significantly affect the resulting speech enhancement performance.

- The DOA-informed modification of the PK-MWF implementation, presented in **Ch. 6**, presents some limitations that still need to be addressed. In the DOA estimation stage, the use of the proposed weighting function for computing the SRP map, which was dependent on the UAV's ego-noise spectral characteristics for a specific rotor speed, needs to be further evaluated in comparison to other alternatives, which could possibly provide more robustness in the DOA estimation with respect to variations in rotor speed due to hovering instability. Additionally, a more elaborate fusion of estimated SPP and DOA information within the implementation of the PK-MWF could be investigated, rather than simply selecting one of the methods to be applied to different frequency ranges. The reliability of speech activity and spatial information could be assessed for different time-frequency frames, such that they can be combined in a more effective manner, and consequently contribute to improvements in not only SNR, but also speech intelligibility and quality.

Appendix A

Appendix to Chapter 2

This appendix is based on a published, open-access international peer-reviewed conference paper [121].

A.1 DOA estimation comparison with MUSIC

In order to evaluate the DOA estimation performance of the GL-LS$_0$ method, i.e., the proposed GL-LS method *without* the additional nonnegativity constraint on the error term (cf. Sec. 2.5.3), two point sources of white Gaussian noise of equal power, denoted σ_S^2, are placed in the far field with an angular separation varying from 30° to 180°. Additive spatially white Gaussian noise of power σ_D^2 is also included and the signal-to-noise ratio (SNR), defined as σ_S^2/σ_D^2, is set to 0 dB and 15 dB in two different simulations.

We then simulate a microphone recording of the resulting signal with the microphone oriented towards the directions of 0°, 60°, 120°, 180°, 240°, and 300°, resulting in $N = 6$ different observation frames. The duration of each frame is 500 ms. The sampling frequency is 16 kHz, and the microphone remains static during one time frame. The PSD $\phi_Y(k, n)$ is estimated using Welch's method, while employing a 512-point Hann window and 50% overlap. The angular dictionary for building the matrix $\widetilde{\mathbf{A}}$ presents a 5° resolution, resulting in a grid with $L = 72$ candidate directions for the point sources. In all simulations, the true DOAs are chosen to be on grid. The group-sparse optimization problem is solved using CVX [159], with the regularization parameter λ in (2.17a) set to $0.01\|\widetilde{\mathbf{A}}^\top \widetilde{\boldsymbol{\varphi}}_Y\|_\infty$, with $\|\cdot\|_\infty$ as the l_∞-norm, and a total of 100 realizations

simulated for each of the scenarios previously described. We evaluate the performance of the proposed DOA estimation method in terms of root-mean-square error (RMSE), computed as:

$$\text{RMSE} = \sqrt{\frac{1}{N_r}\sum_{l=1}^{N_r}\frac{\sum_{p=1}^{P}(\theta_{S_p} - \hat{\theta}^l_{S_p})^2}{P}} \tag{A.1}$$

where N_r is the total number of realizations, θ_{S_p} is the true DOA of source S_p, and $\hat{\theta}^l_{S_p}$ is the estimated DOA of source S_p in realization l.

We compare the RMSE obtained when using the simulated cardioid microphone of flat frequency response with the wideband MUSIC algorithm [20, 21] with harmonic averaging [255] applied to a simulated 6-element uniform linear array (ULA) and to a simulated 6-element uniform circular array (UCA), with a microphone spacing of 5 cm.

The performance comparison between the proposed method for a cardioid microphone with a flat frequency response and the MUSIC algorithm applied to the linear and circular arrays in terms of RMSE for DOA estimation is presented in Fig. A.1. We observe that the proposed method presents overall a lower error than MUSIC for both array geometries and SNR levels under the condition of sufficient angular separation between sources, which situates between 60° and 90° for 0 dB SNR and between 30° and 60° for 15 dB SNR.

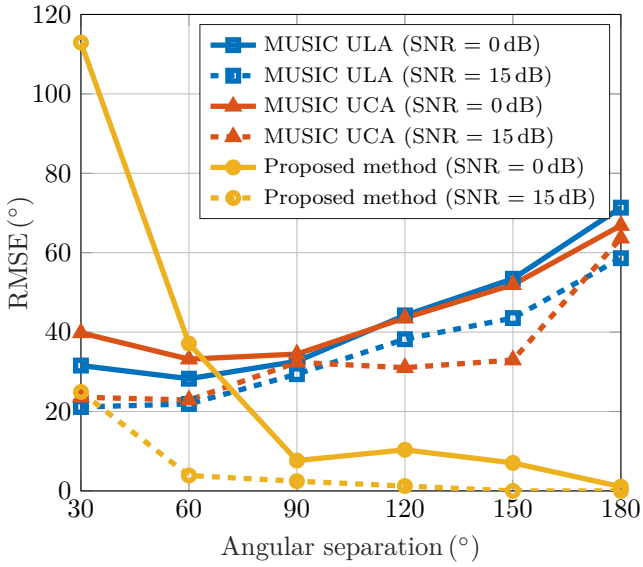


Figure A.1: RMSE of DOA estimation of two stationary point source signals for varying angular separations and different SNRs, obtained with wideband MUSIC and the proposed method.

Appendix B

Appendix to Chapter 3

B.1 GSPICE applied to microphone cross-correlations

Let $\hat{\boldsymbol{\psi}} \in \mathbb{C}^{KP}$ denote the vector of cross-correlations estimated from the microphone array's signals, which is here modeled as

$$\hat{\boldsymbol{\psi}} = \mathbf{A}\boldsymbol{\varphi}_S + \mathbf{e}, \tag{B.1}$$

where $\mathbf{A}$ and $\boldsymbol{\varphi}_S$ are defined as in Sec. 3.4, and $\mathbf{e} \in \mathbb{C}^{KP}$ is the noise component in $\hat{\boldsymbol{\psi}}$. A covariance matrix of the measured cross-correlations is modeled as

$$\mathbf{R} = \tilde{\mathbf{A}}\mathbf{P}\tilde{\mathbf{A}}^H \in \mathbb{R}^{KP\times KP}, \tag{B.2}$$

where $\tilde{\mathbf{A}}$ is an augmented version of $\mathbf{A}$, defined as $\tilde{\mathbf{A}} = [\mathbf{A}, \mathbf{I}_{KP}] \in \mathbb{C}^{KP\times KL+KP}$. The matrix $\mathbf{P}$ is defined as $\mathbf{P} = \mathrm{diag}(\mathbf{p})$, with $\mathbf{p}$ denoting the vector of unknown diagonal loadings within the covariance model used:

$$\mathbf{p} = [\mathbf{p}_1^\top, \ldots, \mathbf{p}_L^\top, \sigma_1, \ldots, \sigma_{KP}]^\top, \tag{B.3}$$

with $\mathbf{p}_l$ corresponding to the l-th group's diagonal loadings vector:

$$\mathbf{p}_l = [p_{1,l}, \ldots, p_{K,l}]^\top, \tag{B.4}$$

and $\sigma_1, \ldots, \sigma_{KP}$ corresponding to the noise power for each frequency and microphone pair, represented as a total of KP single-element groups.

In the GSPICE method, it is first proposed to solve the following optimization problem [200]:

$$\begin{aligned} \underset{\mathbf{p}}{\text{minimize}} \;\; & \hat{\boldsymbol{\psi}}^H \mathbf{R}^{-1} \hat{\boldsymbol{\psi}} + \sum_{l=1}^{L+KP} \nu_l \left\| \mathbf{p}_l \right\|_p \\ \text{subject to} \;\; & \mathbf{R} = \tilde{\mathbf{A}} \mathbf{P} \tilde{\mathbf{A}}^H \\ & \mathbf{p} \geq 0 \end{aligned} \tag{B.5}$$

where ν_l is

$$\nu_l = \left\| \left[\|\tilde{\boldsymbol{a}}_{(l-1)K+1}\|_2^2 \; \ldots \; \|\tilde{\boldsymbol{a}}_{lK}\|_2^2 \right]^\top \right\|_2, \tag{B.6}$$

for $l = 1, \ldots, L$, with $\tilde{\boldsymbol{a}}_q$ denoting the q-th column of $\tilde{\mathbf{A}}$. The objective function in (B.5) corresponds to a relaxed version of the conventional covariance fitting criteria, depending on the choice of the l_p-norm [200], and by introducing the auxiliary variable $\boldsymbol{\beta} = [\boldsymbol{\varphi}_S^\top, \mathbf{e}^\top]^\top$, an equivalent formulation of the optimization problem can be obtained as:

$$\begin{aligned} \underset{\mathbf{p}, \boldsymbol{\beta}}{\text{minimize}} \;\; & \boldsymbol{\beta}^H \mathbf{P}^{-1} \boldsymbol{\beta} + \sum_{l=1}^{L+KP} \nu_l \left\| \mathbf{p}_l \right\|_p \\ \text{subject to} \;\; & \tilde{\mathbf{A}} \boldsymbol{\beta} = \hat{\boldsymbol{\psi}} \\ & \mathbf{p} \geq 0 \end{aligned} \tag{B.7}$$

which, when implemented, can be split into two convex subproblems that are iteratively solved for $\mathbf{p}$ and $\boldsymbol{\beta}$. It is shown in [200] that in the case of homoscedastic noise, i.e., the noise component presents equivariance power, another equivalent formulation of the optimization problem (proof ommitted) can be obtained as:

$$\underset{\boldsymbol{\varphi}_S}{\text{minimize}} \;\; \|\hat{\boldsymbol{\psi}} - \mathbf{A} \boldsymbol{\varphi}_S\|_2 + \sum_{l=1}^{L} \sqrt{\frac{\nu_l}{KP}} \left\| \boldsymbol{\varphi}_{S,l} \right\|_{\frac{2p}{p+1}} \tag{B.8}$$

which, when $p \to \infty$, corresponds to the group square-root Lasso problem *without* a nonnegativity constraint.

Bibliography

[1] A. J. Oxenham, "How we hear: The perception and neural coding of sound," *Annu. Rev. Psychol.*, vol. 69, pp. 27–50, 2018.

[2] N. Kraus and J. Slater, "Beyond words: How humans communicate through sound," *Annu. Rev. Psychol.*, vol. 67, pp. 83–103, 2016.

[3] J. Blauert and J. Allen, *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997.

[4] B. C. Moore, *An Introduction to the Psychology of Hearing*. UK: Brill, 2012.

[5] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annu. Rev. Psychol.*, vol. 42, no. 1, pp. 135–159, 1991.

[6] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, ser. Springer Topics in Signal Processing. Berlin, Heidelberg: Springer, 2008, vol. 1.

[7] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*. Berlin ; London: Springer, 2008.

[8] S. Doclo, S. Gannot, M. Moonen, A. Spriet, S. Haykin, and K. R. Liu, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*. Hoboken, NJ, USA: Wiley, 2010, vol. 9, pp. 269–302.

[9] M. Wölfel and J. McDonough, *Distant Speech Recognition*. UK: John Wiley & Sons Ltd., 2009.

[10] K. Lopatka, J. Kotus, and A. Czyzewski, "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations," *Multimed. Tools. Appl.*, vol. 75, pp. 10 407–10 439, 2016.

[11] S. Krstulović, "Audio event recognition in the smart home," in *Computational Analysis of Sound Scenes and Events*, T. Virtanen, M. D. Plumbley, and D. Ellis, Eds. Cham, Switzerland: Springer, 2018, pp. 335–371.

[12] N. Dey and A. S. Ashour, *Direction of Arrival Estimation and Localization of Multi-Speech Sources*, ser. SpringerBriefs in Electrical and Computer Engineering. Cham, Switzerland: Springer International Publishing, 2018.

[13] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, Providence, RI, USA, 2000.

[14] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. New York, NY, USA: Springer, 2013.

[15] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[16] J. Krolik and D. Swingler, "Multiple broad-band source location using steered covariance matrices," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 10, pp. 1481–1494, 1989.

[17] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 5, no. 3, pp. 288–292, 1997.

[18] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Heidelberg: Springer, 2001, pp. 157–180.

[19] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[20] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[21] B. Friedlander, Ed., *Classical and Modern Direction-of-Arrival Estimation*. Boston: Academic Press, 2009.

[22] R. Roy and T. Kailath, "ESPRIT - Estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, 1989.

[23] A. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *Proc. '83 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '83)*, vol. 8, Boston, MA, USA, 1983, pp. 336–339.

[24] F. Gao and A. Gershman, "A generalized ESPRIT approach to direction-of-arrival estimation," *IEEE Signal Process. Lett.*, vol. 12, no. 3, pp. 254–257, 2005.

[25] M. Rubsamen and A. B. Gershman, "Direction-of-arrival estimation for nonuniform sensor arrays: From manifold separation to fourier domain MUSIC methods," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 588–599, 2009.

[26] M. Pesavento, A. Gershman, and K. Wong, "Direction finding in partly calibrated sensor arrays composed of multiple subarrays," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2103–2115, 2002.

[27] A. B. Gershman, M. Rübsamen, and M. Pesavento, "One-and two-dimensional direction-of-arrival estimation: An overview of search-free techniques," *Signal Process.*, vol. 90, no. 5, pp. 1338–1349, 2010.

[28] G. Su and M. Morf, "The signal subspace approach for multiple wide-band emitter location," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 6, pp. 1502–1522, 1983.

[29] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband MUSIC: Opportunities and challenges for multiple source localization," in *Proc. 2007 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA 2007)*, New Paltz, NY, USA, 2007, pp. 18–21.

[30] B. Ottersten and T. Kailath, "Direction-of-arrival estimation for wide-band signals using the ESPRIT algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 2, pp. 317–327, 1990.

[31] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 4, pp. 823–831, 1985.

[32] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace pseudointensity vectors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 178–192, 2017.

[33] P. Stoica and K. Sharman, "Maximum likelihood methods for direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 7, pp. 1132–1143, 1990.

[34] P.-J. Chung, M. Viberg, and J. Yu, "DOA estimation methods and algorithms," in *Academic Press Library in Signal Processing*, ser. Academic Press Library in Signal Processing, A. M. Zoubir, M. Viberg, R. Chellappa, and S. Theodoridis, Eds. Elsevier, 2014, vol. 3, pp. 599–650.

[35] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 10, pp. 1553–1560, 1988.

[36] J.-Y. Lee, R. E. Hudson, and K. Yao, "Acoustic DOA estimation: An approximate maximum likelihood approach," *IEEE Syst. J.*, vol. 8, no. 1, pp. 131–141, 2014.

[37] Y. Hu, J. Lu, and X. Qiu, "Direction of arrival estimation of multiple acoustic sources using a maximum likelihood method in the spherical harmonic domain," *Appl. Acoust.*, vol. 135, pp. 85–90, 2018.

[38] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[39] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *EEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.

[40] Z. Yang, J. Li, P. Stoica, and L. Xie, "Sparse methods for direction-of-arrival estimation," in *Academic Press Library in Signal Processing*, R. Chellappa and S. Theodoridis, Eds. Academic Press, 2018, pp. 509–581.

[41] S. Fortunati, R. Grasso, F. Gini, M. S. Greco, and K. LePage, "Single-snapshot DOA estimation by using Compressed Sensing," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, p. 120, Dec. 2014.

[42] Q. Shen, W. Liu, W. Cui, and S. Wu, "Underdetermined DOA Estimation Under the Compressive Sensing Framework: A Review," *IEEE Access*, vol. 4, pp. 8865–8878, 2016.

[43] J.-J. Fuchs, "On the application of the global matched filter to DOA estimation with uniform circular arrays," *IEEE Trans. Signal Process.*, vol. 49, no. 4, pp. 702–709, 2001.

[44] J. Velasco, D. Pizarro, and J. Macias-Guarasa, "Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints," *Sensors*, vol. 12, no. 10, pp. 13 781–13 812, 2012.

[45] J. Yin and T. Chen, "Direction-of-arrival estimation using a sparse representation of array covariance vectors," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4489–4493, 2011.

[46] P. Stoica, P. Babu, and J. Li, "SPICE: A sparse covariance-based estimation method for array processing," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 629–638, 2011.

[47] Z. Shi, C. Zhou, Y. Gu, N. A. Goodman, and F. Qu, "Source estimation using coprime array: A sparse reconstruction perspective," *IEEE Sensors Journal*, vol. 17, no. 3, pp. 755–765, 2017.

[48] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, 1997.

[49] D. Malioutov, M. Cetin, and A. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, 2005.

[50] Z. Bai, L. Shi, J. R. Jensen, J. Sun, and M. G. Christensen, "Acoustic DOA estimation using space alternating sparse Bayesian learning," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, no. 1, p. 14, Dec. 2021.

[51] Z. Yang, L. Xie, and C. Zhang, "Off-grid direction of arrival estimation using sparse bayesian inference," *IEEE Trans. Signal Process.*, vol. 61, no. 1, pp. 38–43, 2013.

[52] Z. Tan, P. Yang, and A. Nehorai, "Joint sparse recovery method for compressed sensing with structured dictionary mismatches," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 4997–5008, 2014.

[53] Q. Shen, W. Cui, W. Liu, S. Wu, Y. D. Zhang, and M. G. Amin, "Underdetermined wideband DOA estimation of off-grid sources employing the difference co-array concept," *Signal Process.*, vol. 130, pp. 299–304, 2017.

[54] F. Chen, J. Dai, N. Hu, and Z. Ye, "Sparse Bayesian learning for off-grid DOA estimation with nested arrays," *Digit. Signal Process.*, vol. 82, pp. 187–193, 2018.

[55] J. Dai, X. Bao, W. Xu, and C. Chang, "Root sparse bayesian learning for off-grid DOA estimation," *IEEE Signal Process. Lett.*, vol. 24, no. 1, pp. 46–50, 2017.

[56] W. Cui, Q. Shen, W. Liu, and S. Wu, "Low complexity DOA estimation for wideband off-grid sources based on re-focused compressive sensing with dynamic dictionary," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 5, pp. 918–930, 2019.

[57] Z. Tan, Y. C. Eldar, and A. Nehorai, "Direction of arrival estimation using co-prime arrays: A super resolution viewpoint," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5565–5576, 2014.

[58] Z. Yang and L. Xie, "On gridless sparse methods for multi-snapshot direction of arrival estimation," *Circuits Syst. Signal Process.*, vol. 36, pp. 3370–3384, 2017.

[59] Y. Park and P. Gerstoft, "Gridless sparse covariance-based beamforming via alternating projections including co-prime arrays," *J. Acoust. Soc. Am.*, vol. 151, no. 6, pp. 3828–3837, June 2022.

[60] E. Hänsler and G. Schmidt, Eds., *Speech and Audio Processing in Adverse Environments*, ser. Springer Series on Signals and Communication Technology. Berlin: Springer, 2008.

[61] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2013.

[62] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*. San Rafael, CA, USA: Morgan & Claypool, Jan. 2013, vol. 9.

[63] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[64] B. Van Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[65] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, 2002.

[66] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Process.*, vol. 84, no. 12, pp. 2367–2387, 2004.

[67] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7-8, pp. 636–656, 2007.

[68] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260–276, 2010.

[69] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[70] S. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 478–482, 2000.

[71] T. Gerkmann, M. Krawczyk, and R. Martin, "Speech presence probability estimation based on temporal cepstrum smoothing," in *Proc. 2010 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2010)*, Dallas, TX, USA, 2010, pp. 4254–4257.

[72] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proc. 2011 IEEE Workshop Appl. Signal Process. Audio, Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2011, pp. 145–148.

[73] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 1072–1077, 2010.

[74] Z.-H. Fu, F. Fan, and J.-D. Huang, "Dual-microphone noise reduction for mobile phone application," in *Proc. 2013 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2013)*, Vancouver, BC, Canada, 2013, pp. 7239–7243.

[75] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, "Incorporating the conditional speech presence probability in multi-channel Wiener filter based noise reduction in hearing aids," *EURASIP J. Adv. Signal Process.*, vol. 2009, pp. 1–11, 2009.

[76] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, 2009.

[77] Z. Koldovský, J. Málek, and S. Gannot, "Spatial source subtraction based on incomplete measurements of relative transfer function," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 8, pp. 1335–1347, 2015.

[78] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-Like structure employing eigenvector-based transfer function ratios estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 206–219, 2011.

[79] R. Varzandeh, M. Taseska, and E. A. P. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *Proc. 2017 Hands-Free Speech Commun. Microphone Arrays (HSCMA 2017)*, San Francisco, CA, USA, 2017, pp. 11–15.

[80] M. Taseska and E. A. P. Habets, "DOA-informed source extraction in the presence of competing talkers and background noise," *EURASIP J. Adv. Signal Process.*, vol. 2017, no. 1, p. 60, Dec. 2017.

[81] R. Ali, T. van Waterschoot, and M. Moonen, "An integrated MVDR beamformer for speech enhancement using a local microphone array and external microphones," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, pp. 1–20, 2021.

[82] M. Zohourian, G. Enzner, and R. Martin, "Binaural speaker localization integrated into an adaptive beamformer for hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 515–528, 2018.

[83] R. Van Rompaey and M. Moonen, "Distributed adaptive node-specific signal estimation in a wireless sensor network with partial prior knowledge of the desired source steering vector," in *Proc. 2019 27th European Signal Process. Conf. (EUSIPCO)*, A Coruña, Spain, 2019, pp. 1–5.

[84] O. Thiergart, M. Taseska, and E. A. P. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2182–2196, 2014.

[85] A. Kuklasinski and J. Jensen, "Multichannel Wiener filters in binaural and bilateral hearing aids—speech intelligibility improvement and robustness to DOA errors," *J. Audio Eng. Soc.*, vol. 65, no. 1/2, pp. 8–16, 2017.

[86] I. Cohen, J. Benesty, and S. Gannot, *Speech Processing in Modern Communication: Challenges and Perspectives*. Berlin, Heidelberg: Springer Science & Business Media, 2010.

[87] AW. Bronkhorst and R. Plomp, "Binaural speech intelligibility in noise for hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 86, no. 4, pp. 1374–1383, 1989.

[88] T. Brand, A. Warzybok, J. Rennies, and C. Hauth, "How do humans benefit from binaural listening when recognizing speech in noisy and reverberant conditions?" in *Proc. AES 60th Int. Conf. Dereverberation and Reverberation of Audio, Music, and Speech*, Leuven, Belgium, 2016.

[89] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed sound source localization using relative transfer functions for hearing aid applications," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 611–623, 2017.

[90] S. Launer, J. A. Zakis, and B. C. Moore, "Hearing aid signal processing," in *Hearing Aids*. Springer, 2016, pp. 93–130.

[91] M. F. Dorman, S. Natale, and L. Loiselle, "Speech understanding and sound source localization by cochlear implant listeners using a pinna-effect imitating microphone and an adaptive beamformer," *J. Am. Acad. Audiol.*, vol. 29, no. 03, pp. 197–205, 2018.

[92] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust Automatic Speech Recognition: A Bridge to Practical Applications*. Academic Press, 2015.

[93] T. Kuhn, A. Jameel, M. Stumpfle, and A. Haddadi, "Hybrid in-car speech recognition for mobile multimedia applications," in *Proc. 49th IEEE Veh. Technol. Conf. (VTC '99)*, vol. 3, Houston, TX, USA, 1999, pp. 2009–2013.

[94] M. C. McCallum, J. L. Campbell, J. B. Richman, J. L. Brown, and E. Wiese, "Speech recognition and in-vehicle telematics devices: Potential reductions in driver distraction," *Int. J. Speech Technol.*, vol. 7, no. 1, pp. 25–33, 2004.

[95] B. Lecouteux, M. Vacher, and F. Portet, "Distant speech recognition in a smart home: Comparison of several multisource ASRs in realistic conditions," in *Proc. 12th Annual Conf. Intl. Speech Comm. Assoc. (INTERSPEECH 2011)*, Florence, Italy, 2011, pp. 2273–2276.

[96] A. A. Arriany and M. S. Musbah, "Applying voice recognition technology for smart home networks," in *Proc. 2016 Int. Conf. Eng. MIS (ICEMIS 2016)*, Agadir, Morocco, 2016, pp. 1–6.

[97] S. Bajpai and D. Radha, "Smart phone as a controlling device for smart home using speech recognition," in *Proc. 2019 Int. Conf. Comm. Signal Process. (ICCSP 2019)*, Chennai, India, 2019, pp. 0701–0705.

[98] I. McGraw, R. Prabhavalkar, R. Alvarez, M. G. Arenas, K. Rao, D. Rybach, O. Alsharif, H. Sak, A. Gruenstein, F. Beaufays, and C. Parada, "Personalized speech recognition on mobile devices," in *Proc. 2016 IEEE*

Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2016), Shanghai, China, 2016, pp. 5955–5959.

[99] L. C. De Silva, C. Morikawa, and I. M. Petra, "State of the art of smart homes," *Eng. Appl. Artif. Intell.*, vol. 25, no. 7, pp. 1313–1321, 2012.

[100] P. Chahuara, F. Portet, and M. Vacher, "Context-aware decision making under uncertainty for voice-based control of smart home," *Expert Syst. Appl.*, vol. 75, pp. 63–79, 2017.

[101] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Rob. Auton. Syst.*, vol. 96, pp. 184–210, 2017.

[102] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *Proc. 2015 IEEE IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2015)*, Brisbane, Australia, 2015, pp. 5610–5614.

[103] V. Tourbabin and B. Rafaely, "Direction of Arrival Estimation Using Microphone Array Processing for Moving Humanoid Robots," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 11, pp. 2046–2058, Nov. 2015.

[104] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *Proc. 2004 IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS 2004)*, vol. 3, 2004, pp. 2123–2128 vol.3.

[105] H. M. Do, M. Pham, W. Sheng, D. Yang, and M. Liu, "RiSH: A robot-integrated smart home for elderly care," *Robot. Auton. Syst.*, vol. 101, pp. 74–92, 2018.

[106] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *Proc. 2004 IEEE Int. Conf. Robotics Automation (ICRA 2004)*, vol. 1, New Orleans, LA, USA, 2004, pp. 1033–1038.

[107] L. Marchegiani and X. Fafoutis, "How well can driverless vehicles hear? An introduction to auditory perception for autonomous and smart vehicles," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 3, pp. 92–105, 2022.

[108] D. Gala, N. Lindsay, and L. Sun, "Three-dimensional sound source localization for unmanned ground vehicles with a self-rotational two-microphone array," in *Proc. 5th Int. Conf. Control Dyn. Syst. Robot. (CDSR 2018)*, vol. 104, Niagara Falls, ON, Canada, 2018, pp. 1–11.

[109] H. Sun, X. Liu, K. Xu, J. Miao, and Q. Luo, "Emergency vehicles audio detection and localization in autonomous driving," *arXiv preprint arXiv:2109.14797*, 2021.

[110] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone audition: Sound source localization using on-board microphones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 508–519, 2022.

[111] Y. Hioka, B. Yen, R. McKay, and M. Kingan, "Clean audio recording using unmanned aerial vehicles," in *Unmanned Aerial Systems*. Elsevier, 2021, pp. 175–202.

[112] L. Wang and A. Cavallaro, "A blind source separation framework for ego-noise reduction on multi-rotor drones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2523–2537, 2020.

[113] J. Benesty, J. Chen, and E. A. Habets, *Speech Enhancement in the STFT Domain*. Springer Science & Business Media, 2011.

[114] N. Kehtarnavaz, S. Parris, and A. Sehgal, *Smartphone-Based Real-Time Digital Signal Processing*. Springer Nature, 2022.

[115] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *EEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, 2015.

[116] A. Bertrand, S. Doclo, S. Gannot, N. Ono, and T. van Waterschoot, "Special issue on wireless acoustic sensor networks and ad hoc microphone arrays," *Signal Process.*, vol. 107, no. C, pp. 1–3, 2015.

[117] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. New York, NY, USA: Springer, 2010, vol. 2.

[118] A. Schmidt, H. W. Löllmann, and W. Kellermann, "Acoustic self-awareness of autonomous systems in a world of sounds," *Proc. IEEE*, vol. 108, no. 7, pp. 1127–1149, 2020.

[119] G. Ince, K. Nakamura, F. Asano, H. Nakajima, and K. Nakadai, "Assessment of general applicability of ego noise estimation," in *Proc. 2011 IEEE Int. Conf. Robot. Autom. (ICRA 2011)*, Shanghai, China, 2011, pp. 3517–3522.

[120] A. Schmidt, H. W. Löllmann, and W. Kellermann, "A novel ego-noise suppression algorithm for acoustic signal enhancement in autonomous

systems," in *Proc. 2018 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2018)*, Calgary, AB, Canada, 2018, pp. 6583–6587.

[121] E. Tengan, M. Taseska, T. Dietzen, and T. van Waterschoot, "Direction-of-arrival and power spectral density estimation using a single directional microphone," in *Proc. 2021 29th European Signal Process. Conf. (EUSIPCO)*, Dublin, Ireland, 2021, pp. 221–225.

[122] S. Boyd, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.

[123] E. Tengan, T. Dietzen, S. Ruiz, M. Alkmim, J. Cardenuto, and T. van Waterschoot, "Replication Data for: Speech enhancement using ego-noise references with a microphone array embedded in an unmanned aerial vehicle," KU Leuven RDR, https://doi.org/10.48804/PZAVUC, 2022.

[124] E. Tengan, T. Dietzen, F. Elvander, and T. van Waterschoot, "Direction-of-arrival and power spectral density estimation using a single directional microphone and group-sparse optimization," *EURASIP J. Audio, Speech, Music Process.*, vol. 2023, no. 38, pp. 1–20, 2023.

[125] E. Tengan, T. Dietzen, F. Elvander, and T. van Waterschoot, "Multi-source direction-of-arrival estimation using steered response power and group-sparse optimization," ESAT-STADIUS Tech. Rep. TR 23-163, KU Leuven, Belgium, submitted for publication, Dec. 2023.

[126] M. Alkmim, J. Cardenuto, E. Tengan, T. Dietzen, T. van Waterschoot, J. Cuenca, L. De Ryck, and W. Desmet, "Drone noise directivity and psychoacoustic evaluation using a hemispherical microphone array," *J. Acoust. Soc. Am.*, vol. 152, no. 5, pp. 2735–2745, 2022.

[127] E. Tengan, T. Dietzen, S. Ruiz, M. Alkmim, J. Cardenuto, and T. van Waterschoot, "Speech enhancement using ego-noise references with a microphone array embedded in an unmanned aerial vehicle," in *Proc. 2022 Int. Congr. Acoust. (ICA 2022)*, Gyeongju, South Korea, 2022, pp. 133–143.

[128] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, *et al.*, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, pp. 1–19, 2016.

[129] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. Hoboken, NJ, USA: Wiley, 2018.

[130] F. Elvander, R. Ali, A. Jakobsson, and T. van Waterschoot, "Offline noise reduction using optimal mass transport induced covariance interpolation," in *Proc. 2019 27th European Signal Process. Conf. (EUSIPCO)*, A Coruña, Spain, 2019, pp. 1–5.

[131] Y. Park, F. Meyer, and P. Gerstoft, "Sequential sparse Bayesian learning for time-varying direction of arrival," *J. Acoust. Soc. Am.*, vol. 149, no. 3, pp. 2089–2099, Mar. 2021.

[132] S. Chakrabarty and E. A. P. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *Proc. 2017 IEEE Workshop Appl. Signal Process. Audio, Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2017, pp. 136–140.

[133] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Am.*, vol. 152, no. 1, pp. 107–151, July 2022.

[134] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Process.*, vol. 86, no. 6, pp. 1215–1229, June 2006.

[135] ——, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, July 2001.

[136] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[137] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[138] G. Enzner and P. Thune, "Robust MMSE filtering for single-microphone speech enhancement," in *Proc. 2017 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 4009–4013.

[139] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[140] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[141] N. Fan, J. Rosca, and R. Balan, "Speech Noise Estimation using Enhanced Minima Controlled Recursive Averaging," in *Proc. 2007 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Honolulu, HI, USA, Apr. 2007, pp. IV–581–IV–584.

[142] Jong-Mo Kum and Joon-Hyuk Chang, "Speech Enhancement Based on Minima Controlled Recursive Averaging Incorporating Second-Order Conditional MAP Criterion," *IEEE Signal Process. Lett.*, vol. 16, no. 7, pp. 624–627, July 2009.

[143] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Commun.*, vol. 48, no. 2, pp. 220–231, Feb. 2006.

[144] R. Hendriks, J. Jensen, and R. Heusdens, "Noise Tracking Using DFT Domain Subspace Decompositions," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 3, pp. 541–553, Mar. 2008.

[145] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Square root-based multi-source early PSD estimation and recursive RETF update in reverberant environments by means of the orthogonal Procrustes problem," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 755–769, 2020.

[146] N. Pan, J. Benesty, and J. Chen, "On single-channel noise reduction with rank-deficient noise correlation matrix," *Appl. Acoust.*, vol. 126, pp. 26–35, Nov. 2017.

[147] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: A Deep Learning Approach to MMSE-Based Noise Power Spectral Density Estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1404–1415, 2020.

[148] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional-Recurrent Neural Networks for Speech Enhancement," in *Proc. 2018 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 2401–2405.

[149] X. Li, S. Leglaive, L. Girin, and R. Horaud, "Audio-Noise Power Spectral Density Estimation Using Long Short-Term Memory," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 918–922, June 2019.

[150] M. Kolbk, Z.-H. Tan, and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 153–167, Jan. 2017.

[151] A. Plinge, F. Jacob, R. Haeb-Umbach, and G. A. Fink, "Acoustic Microphone Geometry Calibration: An overview and experimental evaluation of state-of-the-art algorithms," *IEEE Signal Process. Mag.*, vol. 33, no. 4, pp. 14–29, July 2016.

[152] K. Kokkinakis, B. Azimi, Y. Hu, and D. R. Friedland, "Single and multiple microphone noise reduction strategies in cochlear implants," *Trends Amplif.*, vol. 16, no. 2, pp. 102–116, 2012.

[153] A. Saxena and A. Y. Ng, "Learning sound location from a single microphone," in *2009 IEEE Int. Conf. Robot. Autom. (ICRA) Proceedings*, 2009, pp. 1737–1742.

[154] Y. Hioka, R. Drage, T. Boag, and E. Everall, "Direction of arrival estimation using a circularly moving microphone," in *Proc. 2018 Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sept. 2018, pp. 91–95.

[155] S. Jesus, M. Porter, Y. Stephan, X. Demoulin, O. Rodriguez, and E. Coelho, "Single hydrophone source localization," *IEEE J. Oceanic Eng.*, vol. 25, no. 3, pp. 337–346, July 2000.

[156] Q. Shen, W. Liu, W. Cui, S. Wu, Y. D. Zhang, and M. G. Amin, "Low-complexity direction-of-arrival estimation based on wideband co-prime arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1445–1456, 2015.

[157] Y. Hioka and K. Niwa, "PSD estimation in beamspace for source separation in a diffuse noise field," in *Proc. 2014 Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, France, 2014, pp. 85–88.

[158] K. Niwa, T. Kawase, K. Kobayashi, and Y. Hioka, "PSD estimation in beamspace using property of M-matrix," in *Proc. 2016 Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, 2016, pp. 1–5.

[159] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[160] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 68, no. 1, pp. 49–67, Feb. 2006.

[161] B. Porat, *A Course in Digital Signal Processing.* New York, USA: John Wiley, 1997.

[162] Hastie, Trevor, Tibshirani, Robert, and Wainwright, Martin, *Statistical Learning with Sparsity: The Lasso and Generalizations*. New York, USA: CRC Press LLC, 2015.

[163] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted 1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5-6, pp. 877–905, Oct. 2008.

[164] F. Elvander, T. Kronvall, S. Adalbjörnsson, and A. Jakobsson, "An adaptive penalty multi-pitch estimator with self-regularization," *Signal Process.*, vol. 127, pp. 56–70, 2016.

[165] R. Chartrand and Wotao Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proc. 2008 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, USA, Mar. 2008, pp. 3869–3872.

[166] D. Wipf and S. Nagarajan, "Iterative Reweighted l_1 and l_2 Methods for Finding Sparse Solutions," *IEEE J. Sel. Top. Signal Process.*, vol. 4, no. 2, pp. 317–329, Apr. 2010.

[167] J. R. Gilbert, C. Moler, and R. Schreiber, "Sparse matrices in MATLAB: Design and implementation," *SIAM J. Matrix Anal. Appl.*, vol. 13, no. 1, pp. 333–356, 1992.

[168] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices*. Oxford, United Kigndom: Oxford University Press, 2017.

[169] Bang and Olufsen, "Music for Archimedes," CD B&O 101, 1992.

[170] H. Kuttruff, *Room Acoustics*. Boca Raton, FL: CRC Press, 2016.

[171] E. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Tech. Rep., Jan. 2006.

[172] E. De Sena, H. Hacihabiboglu, and Z. Cvetkovic, "On the Design and Implementation of Higher Order Differential Microphones," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 1, pp. 162–174, Jan. 2012.

[173] A. Marti, M. Cobos, and J. J. Lopez, "Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments," in *Proc. 2011 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2011)*, 2011, pp. 2592–2595.

[174] F. Elvander, I. Haasler, A. Jakobsson, and J. Karlsson, "Multi-marginal optimal transport using partial information with applications in robust localization and sensor fusion," *Signal Process.*, vol. 171, p. 107474, 2020.

[175] ——, "Tracking and sensor fusion in direction of arrival estimation using optimal mass transport," in *Proc. 26th European Signal Process. Conf. (EUSIPCO 2018)*, Rome, Italy, 2018, pp. 1617–1621.

[176] ——, "Non-coherent sensor fusion via entropy regularized optimal mass transport," in *Proc. 2019 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2019)*, Brighton, UK, 2019, pp. 4415–4419.

[177] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, 2007.

[178] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?" in *Proc. 2008 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2008)*, Las Vegas, NV, USA, 2008, pp. 2565–2568.

[179] K. D. Donohue, J. Hannemann, and H. G. Dietz, "Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments," *Signal Process.*, vol. 87, no. 7, pp. 1677–1691, 2007.

[180] A. Brutti, M. Omologo, and P. Svaizer, "Multiple source localization based on acoustic map de-emphasis," *EURASIP J. Audio, Speech, Music Process.*, vol. 2010, pp. 1–17, 2010.

[181] M. J. Taghizadeh, P. N. Garner, H. Bourlard, H. R. Abutalebi, and A. Asaei, "An integrated framework for multi-channel multi-source localization and voice activity detection," in *Proc. 2011 Hands-Free Speech Comm. Microphone Arrays (HSCMA 2011)*, Edinburgh, UK, 2011-05, 2011, pp. 92–97.

[182] D. Diaz-Guerra and J. R. Beltran, "Source cancellation in cross-correlation functions for broadband multisource DOA estimation," *Signal Process.*, vol. 170, pp. 1–12, 2020.

[183] Y. Oualil, F. Faubel, D. Klakow, *et al.*, "Joint detection and localization of multiple speakers using a probabilistic interpretation of the steered response power," in *Proc. 2012 Workshop Stat. Perceptual Audition (SAPA-SCALE 2012)*, Portland, OR, USA, 2012, pp. 68–73.

[184] H. Do and H. F. Silverman, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," in *Proc. 2010 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2010)*, Dallas, TX, USA, 2010-03, 2010, pp. 125–128.

[185] ——, "A method for locating multiple sources from a frame of a large-aperture microphone array data without tracking," in *Proc. 2008 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2008)*, Las Vegas, NV, USA, 2008-03, 2008, pp. 301–304.

[186] Y. Oualil, F. Faubel, and D. Klakow, "A fast cumulative steered response power for multiple speaker detection and localization," in *Proc. 21st European Signal Process. Conf. (EUSIPCO 2013)*, Marrakech, Morocco, 2013, pp. 1–5.

[187] M. B. Çöteli, O. Olgun, and H. Hacıhabiboğlu, "Multiple sound source localization with steered response power density and hierarchical grid refinement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2215–2229, 2018.

[188] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2193–2206, 2013.

[189] E. Hadad and S. Gannot, "Multi-Speaker Direction of Arrival Estimation using SRP-PHAT Algorithm with a Weighted Histogram," in *Proc. 2018 IEEE Int. Conf. Science Electrical Eng. Israel (ICSEE 2018)*, Eilat, Israel, 2018, pp. 1–5.

[190] S. Thakallapalli, S. V. Gangashetty, and N. Madhu, "NMF-weighted SRP for multi-speaker direction of arrival estimation: Robustness to spatial aliasing while exploiting sparsity in the atom-time domain," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, no. 1, p. 13, Dec. 2021.

[191] Z. Chu, J. Weng, and Y. Yang, "Determination of propagation model matrix in generalized cross-correlation based inverse model for broadband acoustic source localization," *J. Acoust. Soc. Am.*, vol. 147, no. 4, pp. 2098–2109, 2020.

[192] T. Padois, O. Doutres, F. Sgard, and A. Berry, "Time domain localization technique with sparsity constraint for imaging acoustic sources," *Mech. Syst. Signal Process.*, vol. 94, pp. 85–93, 2017.

[193] D. Bechler and K. Kroschel, "Considering the second peak in the GCC function for multi-source TDOA estimation with a microphone array," in *Proc. 2003 Int. Workshop Acoustic Echo Noise Control (IWAENC 2003)*, Seattle, WA, USA, 2003, pp. 315–318.

[194] E. Tengan, T. Dietzen, F. Elvander, and T. van Waterschoot, "Multi-source direction-of-arrival estimation using group-sparse fitting of steered

response power maps," in *Proc. 2023 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA 2023)*, New Paltz, NY, USA, 2023, pp. 1–5.

[195] G. García-Barrios, J. M. Gutiérrez-Arriola, N. Sáenz-Lechón, V. J. Osma-Ruiz, and R. Fraile, "Analytical model for the relation between signal bandwidth and spatial resolution in steered-response power phase transform (SRP-PHAT) maps," *IEEE Access*, vol. 9, pp. 121 549–121 560, 2021.

[196] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 71–74, 2010.

[197] L. O. Nunes, W. A. Martins, M. V. S. Lima, L. W. P. Biscainho, M. V. M. Costa, F. M. Gonçalves, A. Said, and B. Lee, "A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5171–5183, 2014.

[198] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting a geometrically sampled grid in the steered response power algorithm for localization improvement," *J. Acoust. Soc. Am.*, vol. 141, no. 1, pp. 586–601, 2017.

[199] F. Bunea, J. Lederer, and Y. She, "The group square-root lasso: Theoretical properties and fast algorithms," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 1313–1325, 2014.

[200] T. Kronvall, S. I. Adalbjörnsson, S. Nadig, and A. Jakobsson, "Group-sparse regression using the covariance fitting criterion," *Signal Process.*, vol. 139, pp. 116–130, 2017.

[201] T. Dietzen, E. De Sena, and T. van Waterschoot, "Low-complexity steered response power mapping based on Nyquist-Shannon sampling," in *Proc. 2021 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA 2021)*, New Paltz, NY, USA, 2021, pp. 206–210.

[202] R. J. I. Marks, *Introduction to Shannon Sampling and Interpolation Theory*. Springer, 2012.

[203] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 774–786, 2015.

[204] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. 2015 IEEE*

Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2015), Brisbane, Australia, 2015, pp. 5206–5210.

[205] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," *arXiv preprint arXiv:1612.01840*, 2016.

[206] R. Kapoor, N. Kloet, A. Gardi, A. Mohamed, and R. Sabatini, "Sound propagation modelling for manned and unmanned aircraft noise assessment and mitigation: A review," *Atmosphere*, vol. 12, no. 11, pp. 1–21, 2021.

[207] B. Schäffer, R. Pieren, K. Heutschi, J. M. Wunderli, and S. Becker, "Drone noise emission characteristics and noise effects on humans — A systematic review," *Int. J. Environ. Res. Public Health*, vol. 18, no. 11, p. 5940, 2021.

[208] ISO/FDIS 5305, "(Under development) General requirement of noise measurement of lightweight and small multirotor unmanned aircraft systems (UAS)," Geneva, Switzerland, 2023.

[209] EU, "Regulation (EU) 2018/1139 of the European Parliament and of the Council of 4 July 2018 on Common Rules in the Field of Civil Aviation and Establishing a European Union Aviation Safety Agency, and Amending Regulations (EC) No 2111/2005, (EC) No 1008/2008, (EU) No 996/2010, (EU) No 376/2014 and Directives 2014/30/EU and 2014/53/EU of the European Parliament and of the Council, and Repealing Regulations (EC) No 552/2004 and (EC) No 216/2008 of the European Parliament and of the Council and Council Regulation (EEC) No 3922/91," Aug. 2018.

[210] ——, "Commission Delegated Regulation (EU) 2020/1058 of 27 April 2020 amending Delegated Regulation (EU) 2019/945 as regards the introduction of two new unmanned aircraft systems classes," July 2020.

[211] M. Roger and S. Moreau, "Tonal-noise assessment of quadrotor-type UAV using source-mode expansions," *Acoust.*, vol. 2, no. 3, pp. 674–690, 2020.

[212] H. Jiang, T. Zhou, R. J. Fattah, X. Zhang, and X. Huang, "Multirotor noise scattering by a drone fuselage," in *Proc. 25th AIAA/CEAS Aeroacoustics Conf.*, 2019, p. 2586.

[213] A. D. Thai, E. De Paola, A. Di Marco, L. G. Stoica, R. Camussi, R. Tron, and S. M. Grace, "Experimental and computational aeroacoustic investigation of small rotor interactions in hover," *Appl. Sci.*, vol. 11, no. 10016, pp. 1–26, 2021.

[214] N. Intaratep, W. N. Alexander, W. J. Devenport, S. M. Grace, and A. Dropkin, "Experimental study of quadcopter acoustics and performance at static thrust conditions," in *Proc. 22nd AIAA/CEAS Aeroacoustics Conf.*, Lyon, France, 2016, pp. 1–14.

[215] K. Heutschi, B. Ott, T. Nussbaumer, and P. Wellig, "Synthesis of real world drone signals based on lab recordings," *Acta Acust.*, vol. 4, no. 6, pp. 1–10, 2020.

[216] H. Bian, Q. Tan, S. Zhong, and X. Zhang, "Assessment of UAM and drone noise impact on the environment based on virtual flights," *Aerosp. Sci. Technol.*, vol. 118, p. 106996, 2021.

[217] C. E. Tinney and J. Sirohi, "Multirotor drone noise at static thrust," *AIAA J.*, vol. 56, no. 7, pp. 2816–2826, 2018.

[218] D. Y. Gwak, D. Han, and S. Lee, "Sound quality factors influencing annoyance from hovering UAV," *J. Sound Vib.*, vol. 489, pp. 1–16, 2020.

[219] G. F. Hessler, D. M. Hessler, K. Bay, *et al.*, "Experimental study to determine wind-induced noise and windscreen attenuation effects on microphone response for environmental wind turbine and other applications," *Noise Control Eng. J.*, vol. 56, no. 4, pp. 300–309, 2008.

[220] Y. C. Lemmens, T. Benoit, R. De Roo, and J. Verbeke, "Real-time simulation of an integrated electrical system of a UAV," in *Proc. 2014 SAE Aerosp. Syst. Technol. Conf.*, Cincinnati, OH, USA, Sept. 2014, pp. 1–9.

[221] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "DREGON: Dataset and methods for UAV-embedded sound source localization," in *Proc. 2018 IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS 2018)*, Madrid, Spain, 2018, pp. 1–8.

[222] Y. Hioka, M. Kingan, G. Schmid, and K. A. Stol, "Speech enhancement using a microphone array mounted on an unmanned aerial vehicle," in *Proc. 2016 Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Xi'an, China, 2016, pp. 1–5.

[223] H. H. Hubbard, *Aeroacoustics of Flight Vehicles: Theory and Practice.* NASA Office of Management, Scientific and Technical Information Program, 1991, vol. 1.

[224] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 785–799, 2014.

[225] B. R. Breed and J. Strauss, "A short proof of the equivalence of LCMV and GSC beamforming," *IEEE Signal Process. Lett.*, vol. 9, no. 6, pp. 168–169, 2002.

[226] Y. Hioka, M. Kingan, G. Schmid, R. McKay, and K. A. Stol, "Design of an unmanned aerial vehicle mounted system for quiet audio recording," *Appl. Acoust.*, vol. 155, pp. 423–427, 2019.

[227] C. Veaux, J. Yamagishi, K. MacDonald, *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," p. 15, 2017.

[228] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[229] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. 2001 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2001)*, vol. 2, Salt Lake City, UT, USA, 2001, pp. 749–752.

[230] T. Dietzen, M. Moonen, and T. van Waterschoot, "Instantaneous PSD Estimation for Speech Enhancement based on Generalized Principal Components," in *Proc. 2020 28th European Signal Process. Conf. (EUSIPCO)*, Amsterdam, Netherlands, Jan. 2021, pp. 191–195.

[231] World Health Organization (WHO), "World report on hearing," https://www.who.int/publications/i/item/9789240020481, Mar. 2021.

[232] A. Orji, K. Kamenov, M. Dirac, A. Davis, S. Chadha, and T. Vos, "Global and regional needs, unmet needs and access to hearing aids," *Int. J. Audiol.*, vol. 59, no. 3, pp. 166–172, 2020.

[233] Global Market Insights, "Hearing aids market - by product (behind-the-ear (BTE), completely-in-the-canal/ invisible-in-canal (CIC/IIC), in-the-canal (ITC), in-the-ear (ITE), receiver in the ear/Receiver in canal (RITE/RIC)), by patient (adult, pediatric) and global forecast, 2023 - 2032," https://www.gminsights.com/industry-analysis/hearing-aids-market, Apr. 2023.

[234] ——, "Car audio market size by sales channel (OEM, aftermarket), by component (speaker, amplifier, DSP, microphone, tuner), by sound management (voice recognized, manual), by vehicle (passenger cars, commercial vehicle), by make, forecast 2023 - 2032," https://www.gminsights.com/industry-analysis/car-audio-market, Mar. 2023.

[235] P. Chiariotti, G. Battista, M. Ettorre, and P. Castellini, "Average acoustic beamforming in car cabins: An automatic system for acoustic mapping over 3D surfaces," *Appl. Acoust.*, vol. 129, pp. 47–63, 2018.

[236] K. Nakadai, T. Mizumoto, and K. Nakamura, "Robot-audition-based human-machine interface for a car," in *Proc. 2015 IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS 2015)*, Hamburg, Germany, 2015, pp. 6129–6136.

[237] T. Matheja, M. Buck, and T. Fingscheidt, "A dynamic multi-channel speech enhancement system for distributed microphones in a car environment," *EURASIP J. Adv. Signal Process.*, vol. 2013, pp. 1–21, 2013.

[238] J. Yang, A. Barde, and M. Billinghurst, "Audio augmented reality: A systematic review of technologies, applications, and future research directions," *J. Audio Eng. Soc.*, vol. 70, no. 10, pp. 788–809, 2022.

[239] J. Y. Hong, J. He, B. Lam, R. Gupta, and W.-S. Gan, "Spatial audio for soundscape design: Recording and reproduction," *Appl. Sci.*, vol. 7, no. 6, p. 627, 2017.

[240] A. Evangelista, L. Ardito, A. Boccaccio, M. Fiorentino, A. M. Petruzzelli, and A. E. Uva, "Unveiling the technological trends of augmented reality: A patent analysis," *Comput. Ind.*, vol. 118, pp. 1–15, 2020.

[241] Global Market Insights, "3D audio market - by component (hardware [Speakers, headphones, soundbars, amplifiers] software, services), by end use (commercial [Theaters/Cinemas, automotive, music studios, Events/Concerts], residential [Gaming, AR/VR, mobile devices]), forecast 2023 - 2032," https://www.gminsights.com/industry-analysis/3d-audio-market, May 2023.

[242] M. Wald, "Learning through multimedia: Speech recognition enhancing accessibility and interaction," *J. Educ. Multimed. Hypermedia*, vol. 17, no. 2, pp. 215–233, 2008.

[243] D. Freitas, "Accessibility and design for all solutions through speech technology," in *Speech Technology: Theory and Applications*, F. Chen and K. Jokinen, Eds. New York, NY, USA: Springer US, 2010, pp. 271–299.

[244] K. Hermus, P. Wambacq, and H. Van Hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP J. Adv. Signal Process.*, vol. 2007, pp. 1–15, 2006.

[245] Fortune Business Insights, "Speech and voice recognition market size, share and COVID-19 impact analysis, by technology (voice recognition and speech recognition), by deployment (cloud and on-premise), by end-user (healthcare, IT and telecommunications, automotive, BFSI, government and legal, education, retail and ecommerce,

media and entertainment, and others), and regional forecast, 2023-2030," https://www.fortunebusinessinsights.com/industry-reports/speech-and-voice-recognition-market-101382, June 2023.

[246] ——, "Smart home market size, share and COVID-19 impact analysis, by device type (safety and security devices, energy and water control, climate control, lighting control, consumer electronics), by housing type (multifamily dwelling, single family dwelling), and regional forecast, 2023-2030," https://www.fortunebusinessinsights.com/industry-reports/smart-home-market-101900, July 2023.

[247] S. Luebstorf, J. A. Allen, E. Eden, W. S. Kramer, R. Reiter-Palmon, and N. Lehmann-Willenbrock, "Digging into "zoom fatigue": A qualitative exploration of remote work challenges and virtual meeting stressors," *Merits*, vol. 3, no. 1, pp. 151–166, 2023.

[248] Global Market Insights, "Video conferencing market size by component (hardware [Multipoint control unit (MCU), codecs, peripheral devices], software [On-Premise, cloud], service [Professional, managed]), by type (room-based, telepresence, desktop), application and forecast, 2023-2032," https://www.gminsights.com/industry-analysis/video-conferencing-market, Dec. 2022.

[249] B. Custers, *The Future of Drone Use*. Berlin, Heidelberg: Springer, 2016.

[250] K. Hoshiba, K. Washizaki, M. Wakabayashi, T. Ishiki, M. Kumon, Y. Bando, D. Gabriel, K. Nakadai, and H. G. Okuno, "Design of UAV-embedded microphone array system for sound source localization in outdoor environments," *Sensors*, vol. 17, no. 11, p. 2535, 2017.

[251] Verified Market Research, "Drone services market size and forecast," https://www.verifiedmarketresearch.com/product/drone-services-market, Sept. 2023.

[252] Z. Zheng, Y. Huang, W.-Q. Wang, and H. C. So, "Augmented covariance matrix reconstruction for DOA estimation using difference coarray," *IEEE Trans. Signal Process.*, vol. 69, pp. 5345–5358, 2021.

[253] C. Zhou, Y. Gu, X. Fan, Z. Shi, G. Mao, and Y. D. Zhang, "Direction-of-Arrival Estimation for Coprime Array via Virtual Array Interpolation," *IEEE Trans. Signal Process.*, vol. 66, no. 22, pp. 5956–5971, Nov. 2018.

[254] C. Zhou, Y. Gu, S. He, and Z. Shi, "A robust and efficient algorithm for coprime array adaptive beamforming," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1099–1112, 2018.

[255] M. R. Azimi-Sadjadi, A. Pezeshki, and N. Roseveare, "Wideband DOA estimation algorithms for multiple moving sources using unattended acoustic sensors," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 44, no. 4, pp. 1585–1599, 2008.

Curriculum Vitae



Elisa Tengan received a B.Sc. degree in Electrical Engineering at the Polytechnic School of the University of São Paulo, Brazil, and an M.Sc. degree in Electronics Engineering and Signal Processing at Grenoble INP, France, in 2017. Since 2019, she has been a Ph.D. candidate at KU Leuven, Belgium, in the research group of STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics. Her research focuses on the development of audio signal processing algorithms for sound source localization with constrained microphone setups, as well as for speech enhancement in adverse acoustic conditions. She has served as a reviewer for IEEE/ACM Transactions on Audio, Speech, and Language Processing.

List of publications

International journal publications

1. **E. Tengan**, T. Dietzen, F. Elvander and T. van Waterschoot, "Direction-of-arrival and power spectral density estimation using a single directional microphone and group-sparse optimization," *EURASIP J. Audio, Speech, Music Process.*, vol. 2023, no. 1, p. 38, 2023.

2. M. Alkmim, J. Cardenuto, **E. Tengan**, T. Dietzen, T. van Waterschoot, J. Cuenca, L. De Ryck and W. Desmet, "Drone noise directivity and psychoacoustic evaluation using a hemispherical microphone array," *J. Acoust. Soc. Am.*, vol. 152, no. 5, pp. 2735-2745, 2022.

International conference publications

1. **E. Tengan**, T. Dietzen, F. Elvander and T. van Waterschoot, "Multi-source direction-of-arrival estimation using group-sparse fitting of steered response power maps," in *Proc. 2023 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA 2023)*, New Paltz, NY, USA, Oct. 2023, pp. 1-5.

2. M. Alkmim, J. Cardenuto, **E. Tengan**, T. Dietzen, T. van Waterschoot, J. Cuenca, L. De Ryck and W. Desmet, "Sound quality analysis of a drone using a hemispherical microphone array," in *Proc. 2023 Int. Conf. Noise Vib. Eng. (ISMA-USD)*, Nov. 2022, pp. 4179-4186.

3. **E. Tengan**, T. Dietzen, S. Ruiz, M. Alkmim, J. Cardenuto and T. van Waterschoot, "Speech enhancement using ego-noise references with a microphone array embedded in an unmanned aerial vehicle," in *Proc. 2022 Int. Congr. Acoust. (ICA 2022)*, Gyeongju, South Korea, Oct. 2022, pp. 133-143.

4. **E. Tengan**, M. Taseska, T. Dietzen and T. van Waterschoot, "Direction-of-arrival and power spectral density estimation using a single directional microphone," in *Proc. 2021 29th European Signal Process. Conf. (EUSIPCO 2021)*, Dublin, Ireland, Aug. 2021, pp. 221-225.

Internal reports

1. **E. Tengan**, T. Dietzen, F. Elvander and T. van Waterschoot, "Multi-source direction-of-arrival estimation using steered response power and group-sparse optimization", ESAT-STADIUS Tech. Rep. TR 23-163, KU Leuven, Belgium, submitted for publication, Dec. 2023.

Datasets

1. **E. Tengan**, T. Dietzen, S. Ruiz, M. Alkmim, J. Cardenuto and T. van Waterschoot, "Replication data for: Speech enhancement using ego-noise references with a microphone array embedded in an unmanned aerial vehicle," KU Leuven RDR, `https://doi.org/10.48804/PZAVUC`, 2022.

Acknowledgements

This research work presented in this thesis was carried out at the ESAT Laboratory of KU Leuven, in the frame of FWO Mandate SB 1S86520N, and VLAIO O&O Project no. HBC.2017.0358 “SPOTT - Tomorrow’s Scalable and PersOnalised advertising Technnology, Today”, and KU Leuven internal funds C14/21/075 “A holistic approach to the design of integrated and distributed digital signal processing algorithms for audio and speech communication devices”. The research leading to these results has received funding from the European Research Council under the European Union’s Horizon 2020 research and innovation program/ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors’ views and the Union is not liable for any use that may be made of the contained information.

FACULTY OF ENGINEERING TECHNOLOGY
DEPARTMENT OF ELECTRICAL ENGINEERING
STADIUS CENTER FOR DYNAMICAL SYSTEMS, SIGNAL PROCESSING, AND DATA ANALYTICS
Kasteelpark Arenberg 10
B-3001 Leuven