Spherical Microphone Array Processing for Acoustic Parameter Estimation and Signal Enhancement

by

Daniel P. Jarrett



A thesis submitted in fulfilment of requirements for the degree of
Doctor of Philosophy of Imperial College London

Communications & Signal Processing Group
Department of Electrical & Electronic Engineering
Imperial College London
2013

Illustration: The *Eigenmike* spherical microphone array. (Photo credit: Emanuël Habets)

Copyright declaration



The copyright of this thesis rests with the author and is made available under a Creative Commons *Attribution Non-Commercial No Derivatives* licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Statement of originality

I declare that this thesis and the research to which it refers are the product of my own work under the guidance and supervision of Dr Emanuël A. P. Habets and my thesis supervisor Dr Patrick A. Naylor. Any ideas or quotations from the work of others, published or otherwise, are fully acknowledged in accordance with standard referencing practice. The material of this thesis has not been accepted for any degree, and has not been concurrently submitted for the award of any other degree.

Abstract

In many distant speech acquisition scenarios, such as hands-free telephony or teleconferencing, the desired speech signal is corrupted by noise and reverberation. This degrades both the speech quality and intelligibility, making communication difficult or even impossible. Speech enhancement techniques seek to mitigate these effects and extract the desired speech signal.

This objective is commonly achieved through the use of microphone arrays, which take advantage of the spatial properties of the sound field in order to reduce noise and reverberation. Spherical microphone arrays, where the microphones are arranged in a spherical configuration, usually mounted on a rigid baffle, are able to analyze the sound field in three dimensions; the captured sound field can then be efficiently described in the spherical harmonic domain (SHD).

In this thesis, a number of novel spherical array processing algorithms are proposed, based in the SHD. In order to comprehensively evaluate these algorithms under a variety of conditions, a method is developed for simulating the acoustic impulse responses between a sound source and microphones positioned on a rigid spherical array placed in a reverberant environment.

The performance of speech enhancement algorithms can often be improved by taking advantage of additional *a priori* information, obtained by estimating various acoustic parameters. Methods for estimating two such parameters, the direction of arrival (DOA) of a source (static or moving) and the signal-to-diffuse energy ratio, are introduced.

Finally, the signals received by a microphone array can be filtered and summed by a beamformer. A tradeoff beamformer is proposed, which achieves a balance between speech distortion and noise reduction. The beamformer weights depend on the noise statistics, which cannot be directly observed and must be estimated. An estimation algorithm is developed for this purpose, exploiting the DOA estimates previously obtained to differentiate between desired and interfering coherent sources.

Acknowledgements

First and foremost, I would like to thank my supervisor, Patrick Naylor, for encouraging me to undertake this PhD. In addition, I am thankful for his support, guidance and optimism throughout my time as a PhD student. I would also like to express my appreciation to Emanuël Habets for his supervision, advice and relentless attention to detail, as well as his willingness to give his time so generously. During my PhD I spent four months at *International Audio Laboratories Erlangen*, and I am grateful to Emanuël for this opportunity. I thank my friends and collaborators Mark Thomas and Nikolay Gaubitch for their help with my research.

In addition, I would like to thank my friends and colleagues in the *Communications & Signal Processing* group, including Felicia, James, Jason, Marco, Sira and Ziggy, and my friends outside the office, in particular Benoît, Letizia, Lise, Marine, Nicolas, Pietro, Quentin and Sophie, for making my time as a PhD student more fun and keeping my spirits high. I would like to offer my special thanks to my parents and sister for enduring this process with me, and for their love and moral support.

Finally, I would like to acknowledge the financial support of the *Engineering and Physical Sciences Research Council* (EPSRC) in the form of a Doctoral Training Grant.

Daniel Jarrett

Contents

Page

Copyright declaration 3

Statement of originality 5

Abstract 7

Acknowledgements 9

Contents 11

List of figures 15

List of tables 17

Nomenclature 19
- Abbreviations . . . 19
- General notation . . . 21
- Operators . . . 21
- Symbols and variables . . . 22

Chapter 1. Introduction 25
- 1.1 Context of work . . . 25
- 1.2 Thesis contributions . . . 27
  - 1.2.1 Research statement . . . 27
  - 1.2.2 Publications . . . 28
  - 1.2.3 Original contributions . . . 30
- 1.3 Thesis outline . . . 32

Chapter 2. Background 35
- 2.1 Coordinate systems . . . 35

2.2 Spherical harmonics . . . 37
2.3 Spatial sampling and aliasing . . . 38
2.3.1 Sampling schemes . . . 40
2.4 Array configurations . . . 41
2.5 Beamforming . . . 45
2.6 Associated literature . . . 47

Chapter 3. Acoustic impulse response simulation **49**
3.1 Allen & Berkley's image method . . . 51
3.1.1 Green's function . . . 51
3.1.2 Image method . . . 52
3.2 Proposed method in the spherical harmonic domain . . . 53
3.2.1 Green's function . . . 53
3.2.2 Neumann Green's function . . . 54
3.2.3 Scattering model . . . 55
3.2.4 Proposed method . . . 58
3.3 Implementation . . . 60
3.3.1 Truncation error . . . 60
3.3.2 Computational complexity . . . 62
3.3.3 Algorithm summary . . . 64
3.4 Examples and applications . . . 64
3.4.1 Diffuse sound field energy . . . 64
3.4.2 Binaural interaural time and level differences . . . 68
3.4.3 Mouth simulator . . . 73
3.5 Conclusions . . . 76

Chapter 4. Spatial acoustic parameter estimation **77**
4.1 Direction of arrival estimation . . . 78
4.1.1 Spherical harmonics . . . 79
4.1.2 Direction of arrival estimation using the steered response power 79
4.1.3 Direction of arrival estimation using the pseudointensity vector 84
4.1.4 Computational complexity . . . 87
4.1.5 Performance evaluation . . . 88
4.1.6 Conclusion . . . 91
4.2 Source tracking . . . 93
4.2.1 Problem formulation . . . 93
4.2.2 Eigenbeam-based particle velocity vector estimation . . . 96
4.2.3 Adaptive localization algorithm . . . 96

4.2.4 Performance evaluation . . . 98
4.2.5 Conclusion . . . 101
4.3 Diffuseness estimation . . . 102
4.3.1 Problem formulation . . . 103
4.3.2 Signal-to-diffuse ratio estimation using spatial coherence . . . 105
4.3.3 Diffuseness estimation using the pseudointensity vector . . . 110
4.3.4 Performance evaluation . . . 110
4.3.5 Conclusions . . . 115

Chapter 5. Noise reduction **117**
5.1 Signal model . . . 119
5.1.1 Spatial domain signal model . . . 119
5.1.2 Spherical harmonic domain signal model . . . 120
5.1.3 Mode strength compensation . . . 121
5.2 Tradeoff beamformer . . . 122
5.3 Signal statistics estimation . . . 126
5.3.1 Noise PSD matrix estimation . . . 127
5.3.2 Coherence vector estimation . . . 128
5.4 Desired speech presence probability estimation . . . 129
5.4.1 Multichannel speech presence probability . . . 130
5.4.2 DOA-based probability . . . 130
5.5 Algorithm summary . . . 133
5.6 Performance evaluation . . . 135
5.6.1 Experimental setup . . . 135
5.6.2 Desired speech presence probability . . . 136
5.6.3 Tradeoff beamformer . . . 138
5.7 Conclusions . . . 145

Chapter 6. Conclusions **147**
6.1 Summary of thesis achievements . . . 147
6.2 Future research directions . . . 149

Bibliography **153**

Appendix A. Impulse response generator for MATLAB **169**
A.1 Documentation . . . 169
A.1.1 Function call . . . 169
A.1.2 Notes . . . 171

A.2 Example 171

Appendix B. Spatial correlation in a diffuse sound field **173**

Appendix C. Relationship between the zero-order eigenbeam and the omnidirectional reference microphone signal **175**

List of figures

1.1 Relationship between the problems addressed in the thesis. 31

2.1 Spherical coordinate system. 36
2.2 Beam patterns of the spherical harmonics up to second order. 39
2.3 The GFal acoustic camera. 42
2.4 Mode strength magnitude. 43
2.5 The em32 Eigenmike spherical microphone array. 44

3.1 Illustration of the image method. 53
3.2 Magnitude of an anechoic rigid sphere transfer function. 57
3.3 Polar plot of the magnitude of an anechoic rigid sphere transfer function. 57
3.4 Illustration of the proposed method. 60
3.5 Errors involved in the truncation of the spherical harmonic decomposition in the Green's function and Neumann Green's function. 62
3.6 Pseudocode for the proposed method. 65
3.7 Reverberant sound field energy on the surface of a rigid sphere. 68
3.8 ITDs as a function of source DOA: simulation & ray model approximation. 70
3.9 ILDs in echoic and anechoic environments with the sphere in the centre of the room and a DOA of 0°. 71
3.10 ILDs in echoic and anechoic environments with the sphere near a room wall and a DOA of 0°. 72
3.11 ILDs in echoic and anechoic environments with the sphere near a room wall and a DOA of 100°. 72
3.12 Illustration of the sound energy radiation pattern for a simple mouth and head model. 75

4.1 Plane-wave decomposition beamformer output as a function of the beamformer order. . . . 83
4.2 Beam pattern of the beam P_x, aligned to the x-axis. . . . 86
4.3 Angular errors for the SRP and pseudointensity vector methods. . . . 90
4.4 Power map obtained with real measurements. . . . 92
4.5 Pseudointensity vector DOA estimates obtained with real measurements. 92
4.6 Error in DOA estimates. . . . 99
4.7 Source positions relative to centre of spherical microphone array. . . . 100
4.8 Reference and estimated source positions as a function of time, for various reverberation times. . . . 101
4.9 Mean diffuseness estimated using the proposed and modified CV methods for two frequencies and SNRs. . . . 112
4.10 Standard deviation of the diffuseness estimates obtained using the proposed and modified CV methods for two frequencies and SNRs. . . . 113
4.11 Mean diffuseness estimated using the proposed and modified CV methods for two time averaging lengths. . . . 114
4.12 Standard deviation of the diffuseness estimates obtained using the proposed and modified CV methods for two frequencies and array orders. 114

5.1 Block diagram of the complete noise reduction algorithm, including the tradeoff beamformer and DOA-based statistics estimation. . . . 134
5.2 DOA estimates obtained using 5 source-array positions with identical true DOA. . . . 138
5.3 Time-frequency plots of opening angles, DOA-based probability, *a posteriori* multichannel SPP, and DSPP. . . . 139
5.4 Performance measures as a function of the input signal to coherent noise ratio. . . . 141
5.5 Sample spectrograms of desired speech signal, received signal, and beamformer output signals. . . . 144

A.1 Sample acoustic impulse response and acoustic transfer function generated using SMIRgen. . . . 172

List of tables

5.1 Tradeoff beamformer performance measures as a function of the tradeoff parameters η and μ', for three different scenarios. 143

Nomenclature

Abbreviations

AIR	acoustic impulse response
ATF	acoustic transfer function
BRIR	binaural room impulse response
CV	coefficient of variation
DSB	delay-and-sum beamformer
DirAC	Directional Audio Coding
DOA	direction of arrival
DSPP	desired speech presence probability
HRIR	head-related impulse response
HRTF	head-related transfer function
iSINR	input signal-to-incoherent-noise ratio
ILD	interaural level difference
ITD	interaural time difference
MVDR	minimum variance distortionless response
PCA	principal component analysis
PDF	probability distribution function
PSD	power spectral density
SHD	spherical harmonic domain
SDR	signal-to-diffuse ratio

SNR	signal-to-noise ratio
SPP	speech presence probability
SRA	statistical room acoustics
SRP	steered response power
STFT	short-time Fourier transform
TDOA	time difference of arrival

General notation

x	Scalar quantity
$\mathbf{x}$	Vector quantity
$\mathbf{X}$	Matrix quantity (except $\mathbf{I}$)
X_{lm}	Eigenbeam of order l and degree m
$\widetilde{X}_{lm}$	Mode strength compensated eigenbeam of order l and degree m
$\widetilde{\mathbf{x}}$	Vector of mode strength compensated eigenbeams

Operators

$x * y$	Linear convolution operator
$[\cdot]^*$	Complex conjugate
$[\cdot]^{\mathrm{T}}$	Vector/matrix transpose
$[\cdot]^{-1}$	Matrix inverse
$[\cdot]^{\mathrm{H}}$	Hermitian transpose
$\lvert\cdot\rvert$	Absolute value
$\lceil\cdot\rceil$	Ceiling operator
$\mathrm{diag}\{\cdot\}$	Diagonal operator
$\mathrm{E}\{\cdot\}$	Mathematical expectation
$\mathrm{E_s}\{\cdot\}$	Spatial expectation
$\Re\{\cdot\}$	Real part of a complex number
$\Im\{\cdot\}$	Imaginary part of a complex number
$\mathrm{tr}\{\cdot\}$	Matrix trace
$f'(x)$	Derivative of f with respect to x

Symbols and variables

b_l	Mode strength of order l
β_{ab}	Room boundary reflection coefficient, $a \in \{x, y, z\}, b \in \{1, 2\}$
c	Speed of sound
$\delta(\cdot)$	Dirac delta function
$\delta_{\cdot}$	Kronecker delta
ϵ	Angular error
f	Frequency
$f(\cdot)$	Probability distribution function
γ	Spatial coherence
$\boldsymbol{\gamma}$	Propagation/coherence vector
Γ	Signal-to-diffuse ratio
g	Time domain free-space Green's function
G	Frequency domain free-space Green's function
G_{N}	Frequency domain Neumann Green's function
$\mathfrak{h}$	Filter weights vector
h	Acoustic impulse response (time domain)
H	Acoustic transfer function (frequency domain)
$h_l^{(1)}$	Spherical Hankel function of the first kind and of order l
i	Complex number, $i^2 = -1$
$\mathcal{I}$	Intensity vector
$\mathbf{I}$	Pseudointensity vector
$\mathbf{I}_{N\times N}$	$N \times N$ identity matrix

j_l	Spherical Bessel function of order l
k	Wavenumber (continuous)
$\dot{k}$	Wavenumber (discrete)
$\dot{\mathfrak{k}}$	Frequency index
λ	Lagrange multiplier
l	Order
ℓ	Time index
L	Array order (maximum spherical harmonic order)
m	Degree
$\mathcal{M}_{\text{ref}}$	Reference microphone (at centre of sphere)
μ	Tradeoff parameter
n	Time (discrete)
$\Omega = (\theta, \phi)$	Spherical coordinates (inclination θ, azimuth ϕ)
Ω_0	Direction of arrival
Ω_{u}	Beamformer look direction
Ψ	Diffuseness
$\mathbf{\Phi}$	Covariance matrix
$\mathcal{P}_l$	Legendre polynomial of order l
$\mathcal{P}_{lm}$	Associated Legendre function (or polynomial) of order l and degree m
q	Microphone index
Q	Number of microphones
$\mathbf{r}$	Receiver position vector
$\mathbf{r}_{\text{s}}$	Source position vector
ρ_0	Density of air

t	Time (continuous)
T_{60}	Reverberation time
$\mathbf{u}$	Unit vector pointing towards acoustic source
$\hat{\mathbf{u}}$	Estimate of $\mathbf{u}$
$\mathbf{v}$	Particle velocity vector
Y_{lm}	Spherical harmonic of order l and degree m
Z	Beamformer output signal

Chapter 1

Introduction

1.1 Context of work

The motivation behind the work presented in this thesis lies in the rapidly growing demand for speech communication systems over the last couple of decades. Such systems are now commonplace in our everyday lives, primarily for human-human communication. However, as the most natural form of human communication, speech also promises to play an ever-growing part in *human-machine communication*. While speech-based interfaces were once confined to the realms of science fiction, they are now becoming an increasingly popular way of interacting with devices such as smartphones, desktop and tablet computers, robots or televisions. This trend has been fueled by advances in speech recognition and synthesis technology, as well as the explosion in available computing power, particularly on mobile devices.

The field of acoustic signal processing seeks to solve a number of problems relating to these systems, which can broadly be divided into two categories: **acoustic parameter estimation** and **acoustic signal enhancement**. Acoustic parameter estimation involves the estimation of parameters such as the location or direction of arrival of one or more acoustic sources, the diffuseness of a sound field, the number of sources present in a sound field, or the reverberation time of an acoustic environment.

In many speech communication systems, the speech to be acquired originates from a *distant* speaker (located far away from the microphone or microphones). While in some applications, such as teleconferencing systems, a close-talking microphone forming part of a headset may be available, in others, such as hearing aids or assistive listening devices, this is a far less practical option. As a result, the acquired speech is corrupted by the surrounding environment. One major cause for this degradation is the presence of noise, where by 'noise' we mean any acoustic signal which is undesired, e.g., an interfering speech signal or background noise. The other is the presence of obstacles to the propagation of sound waves, in particular room boundaries (walls, floors and ceiling), which cause *reverberation*.

These effects degrade the quality of the acquired speech, and in some cases, its intelligibility, making communication difficult or even impossible. Acoustic signal enhancement or speech enhancement techniques seek to mitigate these effects, and extract the desired (usually speech) signal. The main problems of interest within speech enhancement are noise reduction, echo cancellation and dereverberation. Although the speakerphone was first released by AT&T in 1954 [34], these remain unsolved problems.

Acoustic signal processing problems are commonly approached with microphone arrays [11,17,36], i.e., an arrangement of microphones in a specific configuration, thereby taking advantage of the spatial properties of the sound field (or *spatial diversity*) in order to improve performance. Owing to the similarity of the problems involved, many microphone array processing techniques are based on narrowband antenna array processing techniques [25]; however, microphone array processing faces its own unique challenges [11]. These include the broadband nature of speech (which covers several octaves), the non-stationarity of speech, and the fact that the desired and noise signals often have very similar spectral characteristics [11]. In addition, the placement and number of microphones is restricted, primarily by cost, aesthetics and available space. Considerations of space limit both the inter-microphone spacing and total microphone array size, and are of particular importance for devices operating in confined spaces,

such as hearing aids.

In theory, any microphone array configuration is possible; in practice, most microphone arrays are planar, i.e., the microphones lie on a flat, two-dimensional surface. Real sound fields are three-dimensional, however, and can only be properly analyzed with a three-dimensional array. The spherical configuration is convenient due to its symmetry and equal performance in all directions. In addition, the captured sound field can be efficiently described in the spherical harmonic domain [77], based on a formulation of the wave equation in spherical coordinates. Spherical microphone arrays [1, 90] are usually either *open* or *rigid*, i.e., the microphones are either suspended in free space or mounted on a rigid baffle. They have recently started to become commercially available, in the form of products such as the *acoustic camera* by GFal, the *Eigenmike* by mh acoustics, or the *RealSpace Panoramic Audio Camera* by VisiSonics, yet to date there have been few algorithms designed for these arrays. It is in this context that we make the contributions contained in this thesis.

1.2 Thesis contributions

1.2.1 Research statement

The aim of this thesis is to exploit the properties of spherical microphone arrays and the spherical harmonic domain (SHD), and propose acoustic parameter estimation and signal enhancement algorithms that are capable of operating in noisy reverberant environments.

1.2.2 Publications

The following publications were produced during the course of this work:

Journal publications

[J1] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: algorithm and applications," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1462–1472, Sep. 2012.

[J2] D. P. Jarrett and E. A. P. Habets, "On the noise reduction performance of a spherical harmonic domain tradeoff beamformer," *IEEE Signal Process. Lett.*, vol. 19, no. 11, pp. 773–776, Nov. 2012.

[J3] D. P. Jarrett, M. Taseska, E. A. P. Habets and P. A. Naylor, "Noise reduction in the spherical harmonic domain using a tradeoff beamformer and narrowband DOA estimates," submitted to *IEEE Trans. Audio, Speech, Lang. Process.*.

Conference and workshop publications

[C1] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 442–446.

[C2] ——, "Eigenbeam-based acoustic source tracking in noisy reverberant environments," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov. 2010, pp. 576–580.

[C3] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Simulating room impulse responses for spherical microphone arrays," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 129–132.

[C4] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, N. D. Gaubitch, and P. A. Naylor, "Dereverberation performance of rigid and open spherical microphone arrays: theory & simulation," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Edinburgh, UK, Jun. 2011, pp. 145–150.

[C5] D. P. Jarrett, E. A. P. Habets, J. Benesty, and P. A. Naylor, "A tradeoff beamformer for noise reduction in the spherical harmonic domain," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Aachen, Germany, Sep. 2012.

[C6] D. P. Jarrett, O. Thiergart, E. A. P. Habets, and P. A. Naylor, "Coherence-based diffuseness estimation in the spherical harmonic domain," in *Proc. IEEE Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, Eilat, Israel, Nov. 2012.

[C7] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "Spherical harmonic domain noise reduction using an MVDR beamformer and DOA-based second-order statistics estimation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 654–658.

[C8] S. Braun, D. P. Jarrett, J. Fischer, and E. A. P. Habets, "An informed spatial filter for dereverberation in the spherical harmonic domain," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 669–673.

The contributions contained in [J2], [C4] and [C8] are not presented in this thesis.

1.2.3 Original contributions

The following aspects of the thesis are, to the best of the author's knowledge, original contributions:

- Development of a rigid sphere acoustic impulse response estimation method. *(Chapter 3, published in [C3,J1])*
  - Comparison of a theoretical prediction of reverberant sound energy on the surface of a rigid sphere to simulated results obtained using the proposed method. *(Section 3.4.1)*
  - Analysis of interaural time differences and interaural level differences in a reverberant environment using the proposed method. *(Section 3.4.2)*
- Development of a pseudointensity vector–based direction-of-arrival estimation method employing zero- and first-order eigenbeams. *(Section 4.1, published in [C1])*
  - Formulation and implementation of a steered response power–based direction-of-arrival estimation method, and comparison with the proposed method. *(Sections 4.1.2 and 4.1.5)*
- Development of a particle velocity vector–based source tracking method. *(Section 4.2, published in [C2])*
  - Derivation of an adaptive filter for particle velocity estimates. *(Section 4.2.3.1)*
- Development of a diffuseness estimation algorithm based on the coherence between eigenbeams. *(Section 4.3, published in [C6])*
  - Derivation of an expression for the coherence between eigenbeams in a sound field composed of both directional and diffuse components. *(Section 4.3.2.1)*
  - Implementation of a coefficient of variation–based diffuseness estimation algorithm, and comparison with the proposed method. *(Sections 4.3.3 and 4.3.4)*

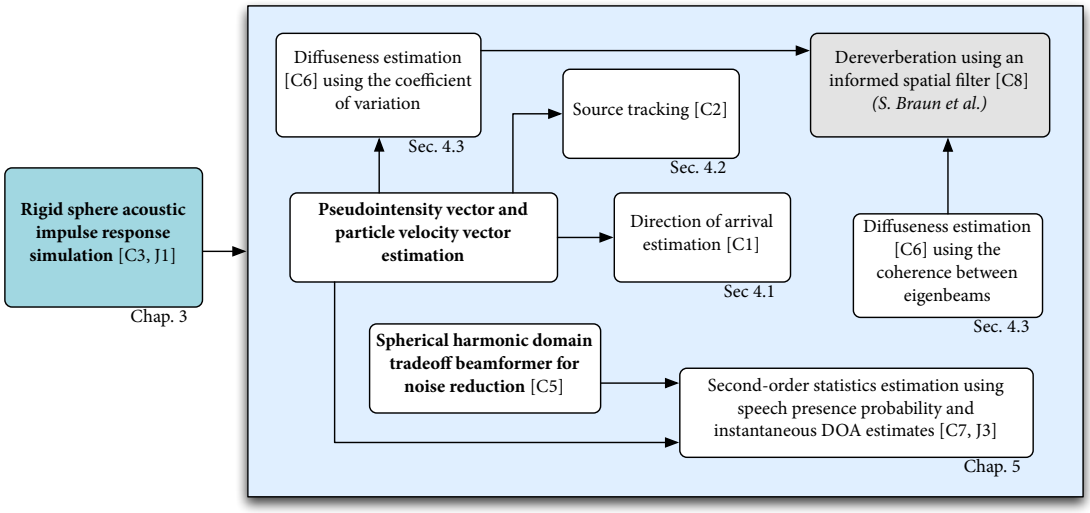


Figure 1.1: Relationship between the problems addressed in the thesis.

- Development of a noise reduction algorithm to suppress both spatially coherent and spatially incoherent noise. *(Chapter 5, published in/submitted to [C7,J3])*
  - Formulation and implementation of a SHD tradeoff beamformer. *(Section 5.2, published in [C5])*
  - Development of a signal statistics estimation algorithm, which is necessary to compute the weights of the tradeoff beamformer. *(Section 5.3)*
  - Evaluation of the proposed noise reduction algorithm using measured acoustic impulse responses. *(Section 5.6)*

The relationship between each of the problems addressed in this thesis is summarized in Fig. 1.1, which also indicates which publications and thesis sections/chapters relate to each problem.

1.3 Thesis outline

The content of this thesis is structured as follows:

- In **Chapter** 2, the fundamentals of spherical array processing are reviewed. This includes an introduction to spherical harmonics, spatial sampling and aliasing, spherical array configurations, and some simple beamforming techniques.

- In **Chapter** 3, a method is proposed for simulating the acoustic impulse responses between a sound source and the microphones positioned on a spherical array, taking into account specular reflections of the source by employing the well-known image method, and scattering from the rigid sphere by employing spherical harmonic decomposition. This method is necessary to comprehensively evaluate spherical array processing algorithms under many acoustic conditions. Three examples are presented: an analysis of a diffuse reverberant sound field, a study of binaural cues in the presence of reverberation, and an illustration of the algorithm's use as a mouth simulator.

- **Chapter** 4 presents novel parameter estimation algorithms in the SHD. We first propose a low-complexity method for direction of arrival estimation based on a pseudointensity vector, and compare it to a steered response power localization method. We then propose an adaptive source tracking algorithm, where the tracking is performed using an adaptive principal component analysis of the particle velocity vector. The pseudointensity and particle velocity vectors are estimated using a spherical microphone array, and are formed by combining the zero- and first-order eigenbeams, which result from a spherical harmonic decomposition of the sound field. Finally, we propose a diffuseness estimator based on the coherence between eigenbeams. The weighted averaging of the diffuseness estimates over all eigenbeam pairs, unlike in the spatial domain where the diffuseness is typically estimated using the coherence between a pair of microphones, is shown

to significantly reduce the variance of the estimates, particularly in fields with low diffuseness.

- In **Chapter 5** we present a tradeoff beamformer in the SHD that enables a tradeoff between noise reduction and speech distortion. This beamformer includes the SHD minimum variance distortionless response (MVDR) and multichannel Wiener filters as special cases. We propose an algorithm to estimate the second-order statistics of the noise and desired signal using a speech presence probability–based method that can distinguish between a coherent desired source and a coherent noise source. We show that the tradeoff beamformer is able to reduce high levels of coherent noise with low speech distortion.

- The thesis is concluded and future work is discussed in **Chapter 6**.

Chapter 2

Background

The sound field captured at a point $\mathbf{r}$ in space and time t is denoted as $p(t, \mathbf{r})$. By applying the temporal Fourier transform to $p(t, \mathbf{r})$, we obtain the sound pressure $P(k, \mathbf{r})$, where k denotes the wavenumber and is related to the angular frequency ω and speed of sound c via the dispersion relation $k = \omega/c$. We assume the acoustic waves propagate in a non-dispersive medium, such that the propagation speed c is independent of the wavenumber k.

2.1 Coordinate systems

Unless otherwise indicated, in this thesis we work in spherical coordinates $\mathbf{r} = (r, \Omega) = (r, \theta, \phi)$, with radial distance r, inclination θ and azimuth ϕ. We adopt the spherical coordinate system used in [35, 76, 112, 119], which is illustrated in Fig. 2.1. The spherical coordinates are related to Cartesian coordinates x, y, z via the expressions [119, eqn. 2.47]

$$x = r \sin\theta \cos\phi \tag{2.1a}$$

$$y = r \sin\theta \sin\phi \tag{2.1b}$$

$$z = r \cos\theta. \tag{2.1c}$$

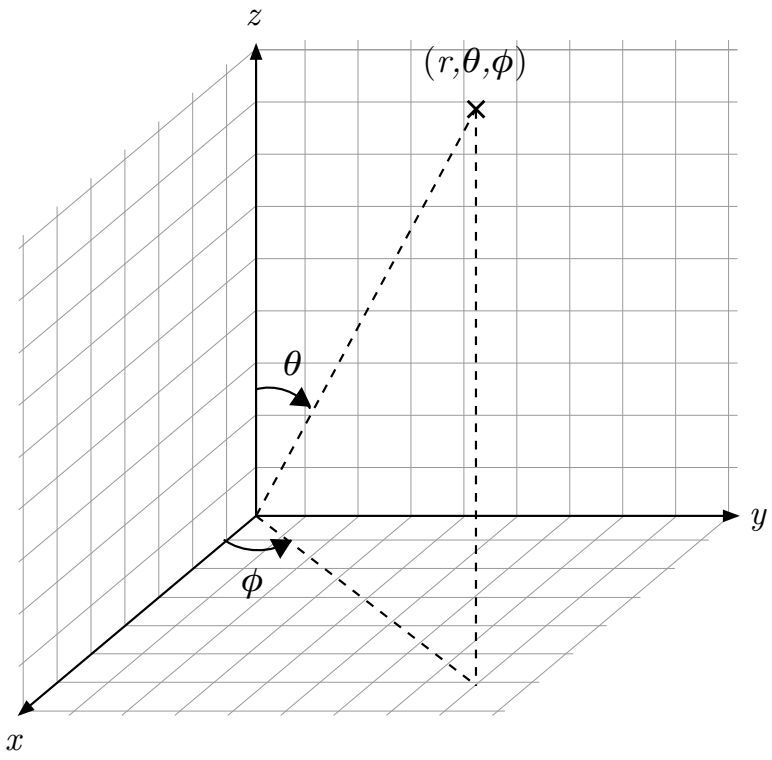


Figure 2.1: Spherical coordinate system used in this thesis, defined relative to Cartesian coordinates. The radial distance r is the distance between the observation point and the origin of the coordinate system. The inclination θ is measured from the positive z-axis, and the azimuth ϕ is measured in the xy-plane from the positive x-axis.

Conversely, the spherical coordinates may be obtained from the Cartesian coordinates using

$$r = \sqrt{x^2 + y^2 + z^2} \tag{2.2a}$$

$$\theta = \arccos\left(\frac{z}{r}\right) \tag{2.2b}$$

$$z = \arctan\left(\frac{y}{x}\right), \tag{2.2c}$$

where arctan is the four-quadrant inverse tangent (implemented using the function `atan2()` in most environments).

2.2 Spherical harmonics

The sound field captured by a spherical array can be conveniently described in the spherical harmonic domain (SHD). The spatial domain signals $P(k, \mathbf{r})$ are expanded into a series of orthogonal basis functions, the spherical harmonics Y_{lm} [119], via the expression [119, eqn. 6.48]

$$P(k, \mathbf{r}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} P_{lm}(k)\, Y_{lm}(\Omega), \tag{2.3}$$

which is referred to as a *spherical harmonic(s) expansion* or *spherical harmonic decomposition* of the sound field. The coefficients $P_{lm}(k)$, which can be considered as counterparts to the Fourier series coefficients in one dimension, are often called *eigenbeams* to reflect the fact that the spherical harmonics are eigensolutions of the wave equation in spherical coordinates [7, 119], and are given by [119, eqn. 6.48]

$$P_{lm}(k) = \int_{\Omega \in \mathcal{S}^2} P(k, \mathbf{r})\, Y_{lm}^*(\Omega) \mathrm{d}\Omega, \tag{2.4}$$

where $\int_{\Omega \in \mathcal{S}^2} \mathrm{d}\Omega = \int_0^{2\pi} \mathrm{d}\phi \int_0^{\pi} \sin\theta \mathrm{d}\theta$ and $(\cdot)^*$ denotes the complex conjugate. The operations in (2.4) and (2.3) are respectively referred to as the forward and inverse *spherical Fourier transform*; the parameters of the spherical Fourier transform are the order l and degree m.

The spherical harmonic of order l and degree m is defined as [119, eqn. 6.20]

$$Y_{lm}(\Omega) = \sqrt{\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!}} \mathcal{P}_{lm}(\cos\theta) e^{im\phi}, \tag{2.5}$$

where $\mathcal{P}_{lm}$ is the associated Legendre function[1] and $i = \sqrt{-1}$. The beam patterns of the

[1] In this thesis, for consistency with spherical array processing literature, we refer to l as the order and m as the degree of the spherical harmonics and associated Legendre functions (or polynomials). However, it should be noted that in other fields, l is referred to as the degree, and m as the order. This reflects the fact that the words *degree* and *order* are used interchangeably when referring to polynomials.

spherical harmonics up to second order are illustrated in Fig. 2.2. It can be seen that the zero-order spherical harmonic is omnidirectional, while the first-order spherical harmonics have a dipole directivity pattern.

The spherical harmonics exhibit a useful property that we will make use of later in this thesis, namely that they are mutually orthonormal [119, eqn. 6.45], i.e.,

Property 2.2.1.

$$\int_{\Omega \epsilon \mathcal{S}^2} Y_{lm}(\Omega) Y_{pq}^*(\Omega) \mathrm{d}\Omega = \delta_{lp}\delta_{mq}, \tag{2.6}$$

where the Kronecker delta δ is defined as follows:

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j; \\ 0, & \text{if } i \neq j. \end{cases} \tag{2.7}$$

2.3 Spatial sampling and aliasing

In practice a continuous pressure sensor is not available, and the sound field must be spatially sampled, such that the integral in (2.4) is replaced by a sum over a discrete number of microphones Q at positions $\mathbf{r}_q$, $q = 1, \ldots Q$ [78, 90, 95]

$$P_{lm}(k) = \int_{\Omega \epsilon \mathcal{S}^2} P(k, \mathbf{r}) Y_{lm}^*(\Omega) \mathrm{d}\Omega \tag{2.8a}$$

$$\approx \sum_{q=1}^{Q} \mathcal{g}_{q,lm} P(k, \mathbf{r}_q). \tag{2.8b}$$

This is a quadrature rule: the approximation of a definite integral by a weighted sum. The quadrature weights $\mathcal{g}_{q,lm}$ are chosen such that the error involved in this approximation is minimized, and are a function of the sampling configuration chosen. Error-free sampling is achieved when the approximation in (2.8b) becomes an equality, or equivalently, when the discrete orthonormality error is zero [90], i.e.,

$$\sum_{q=1}^{Q} \mathcal{g}_{q,lm} Y_{l'm'}(\Omega_q) = \delta_{l-l'}\delta_{m-m'}. \tag{2.9}$$

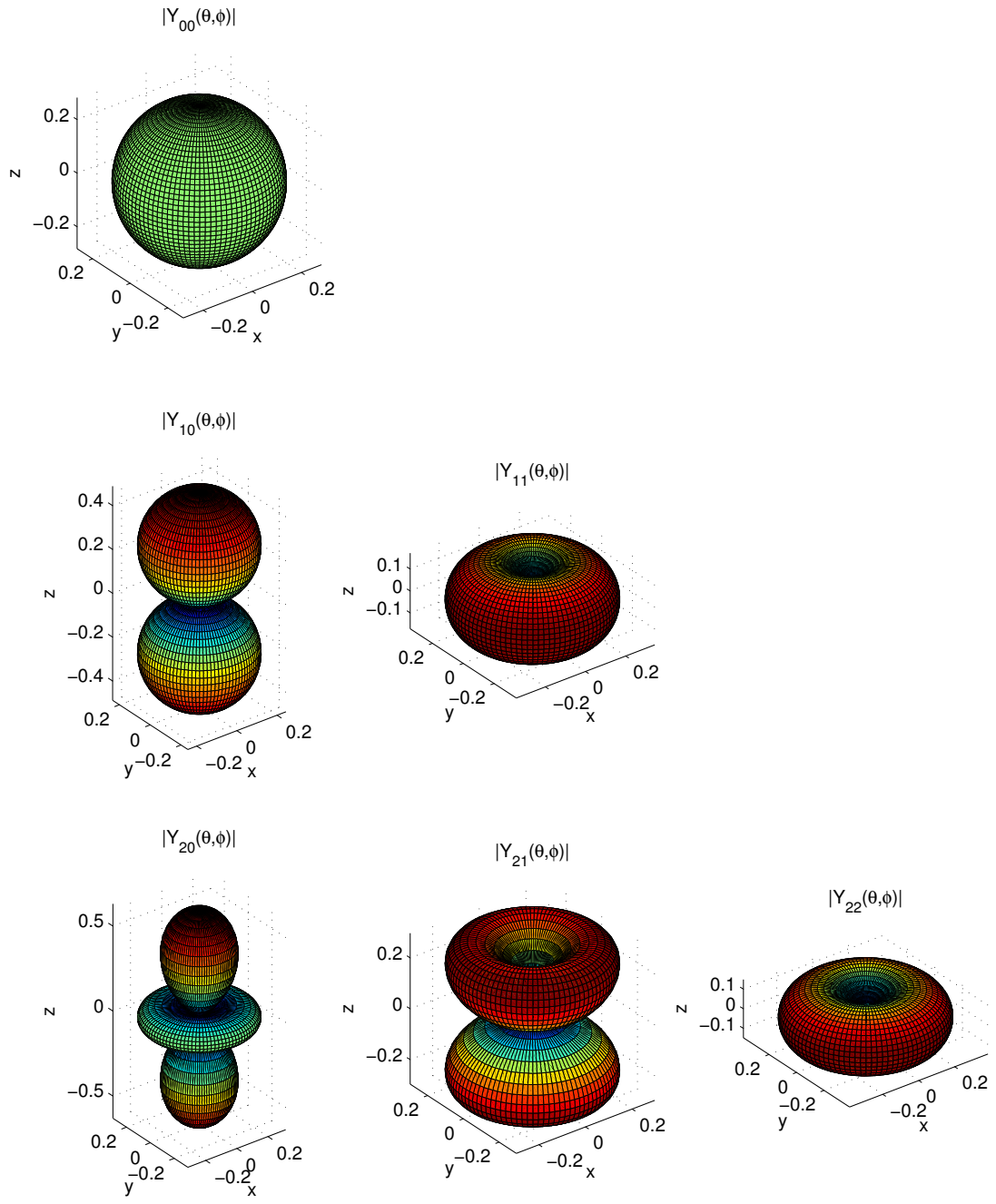


Figure 2.2: Beam patterns $|Y_{lm}(\theta, \phi)|$ of some of the most commonly used spherical harmonics, for $\{l \in \mathbb{Z} | 0 \leq l \leq 2\}$, $\{m \in \mathbb{Z} | 0 \leq m \leq l\}$. The beam patterns for $m < 0$ are omitted as they are identical to those for $m > 0$.

In the same way that a time domain signal must be temporally band-limited in order to be fully reconstructed from a finite number of samples without temporal aliasing, the SHD sound field must be order-limited ($P_{lm} = 0$ for $l > L$, where L is the order of the sound field) to be captured with a finite number of microphones without spatial aliasing [90]. A sound field which is limited to an order L is represented using a total of $\sum_{l=0}^{L}\sum_{m=-l}^{l} 1 = \sum_{l=0}^{L}(2l+1) = (L+1)^2$ eigenbeams, therefore all spatial sampling schemes require at least $(L+1)^2$ microphones to sample a sound field of order L without spatial aliasing.

Spatial aliasing occurs when high-order sound fields are captured using an insufficient number of sensors and the high-order eigenbeams are aliased into the lower orders. A number of sampling schemes, three of which are presented below, are aliasing-free (or have negligible aliasing) for order-limited functions. However, in practice, sound fields are not order-limited: they are represented by an infinite series of spherical harmonics [94]. Nevertheless, the magnitude of the eigenbeams decays rapidly for $l > kr$ (see Section 2.4). We can therefore consider the aliasing error to be negligible if $kr < L$ [90,94], or equivalently if the operating frequency f satisfies $f < \frac{Lc}{2\pi r}$, where c is the speed of sound, and the frequency f and wavenumber k are related via the expression $f = \frac{kc}{2\pi}$. This means that for a sound field of order $L = 4$ and an array radius of $r = 4.2$ cm (the radius of the *Eigenmike* [79]), the operating frequency must be smaller than 5.2 kHz, for example. For higher operating frequencies, Rafaely *et al.* proposed spatial anti-aliasing filters to reduce the aliasing errors [94].

2.3.1 Sampling schemes

The simplest sampling scheme is **equiangle sampling**, where the inclination θ and azimuth ϕ are uniformly sampled at $2(L+1)$ angles given by $\theta_i = \frac{\pi i}{2L+2}$, $i = 0, \ldots, 2L+1$ and $\phi_j = \frac{2\pi j}{2L+2}$, $j = 0, \ldots, 2L+1$ [32,90]. The scheme therefore requires a total of $Q = 4(L+1)^2$ microphones. The quadrature weights are given by $\mathcal{G}_{q,lm} = \mathcal{G}_i Y_{lm}^*(\theta_i, \phi_j)$ [32,90], where $q = j + i(2L+2) + 1$, and the term $\mathcal{G}_i$ compensates for the denser sampling in θ near the

poles [32, 90]. The advantage of this scheme is the uniformity of the angle distributions, which can be useful when samples are taken by a rotating microphone, however this comes at the expense of a relatively large number of required samples.

In **Gaussian sampling**, only half as many samples are needed: the azimuth is still sampled at $2(L+1)$ angles, whereas the inclination is sampled at only $L+1$ angles, requiring a total of $2(L+1)^2$ microphones. The azimuth angles are the same as for equiangle sampling, while the inclination angles must satisfy $\mathcal{P}_{L+1}(\cos\theta_i) = 0$, $i = 0, \ldots, L$ [90], where $\mathcal{P}_{L+1}$ is the Legendre polynomial of order $L+1$. The quadrature weights are then given by $\mathcal{G}_{q,lm} = \mathcal{G}_i Y_{lm}^*(\theta_i, \phi_j)$ [94], where $q = j + i(2L+2) + 1$ and the weights $\mathcal{G}_i$ are given in [7, 67]. The disadvantage of this scheme is that the inclination distribution is no longer uniform, however for a fixed array configuration this is not likely to be a problem.

Finally, in **(quasi) uniform sampling**, the samples are (quasi) uniformly distributed on the sphere, i.e., the distance between each sample and its neighbours is (quasi) constant. A limited number of distributions perfectly satisfy this requirement: the vertices of the so-called platonic solids. However, there are a number of nearly uniform distributions with negligible orthogonality error, which require at least $(L+1)^2$ microphones. The quadrature weights are given by $\mathcal{G}_{q,lm} = \frac{4\pi}{Q} Y_{lm}^*(\Omega_q)$ for uniform sampling [35, 119].

In the rest of this thesis, uniform sampling will be employed, and it will be assumed that this sampling is aliasing-free. This is a reasonable assumption for the operating frequencies (up to 4 kHz) considered in this work.

2.4 Array configurations

The sound pressure captured by the microphones in a spherical array depends on the array properties, e.g., radius, configuration (open, rigid, dual-sphere, etc.), or microphone type. This dependence is captured by the frequency-dependent *mode strength* $b_l(k)$, which determines the amplitude of the l^{th}-order eigenbeam(s) $P_{lm}(k)$ $(m = -l, \ldots, l)$. For a unit amplitude plane wave incident from a direction Ω_0, the SHD sound pressure

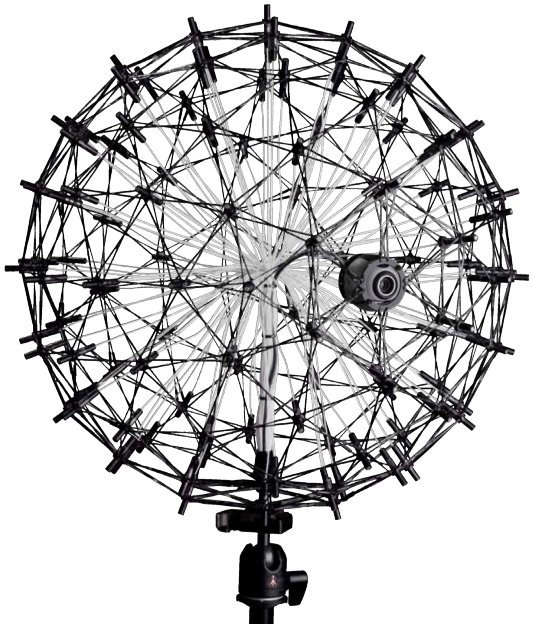


Figure 2.3: The GFal Sphere120 pro acoustic camera. This open array of radius 30 cm is comprised of 120 microphones, as well as a digital camera. © gfai tech GmbH, used with permission.

and the mode strength $b_l(k)$ are related via the expression [77, 89, 112]

$$P_{lm}(k) = b_l(k) Y_{lm}^*(\Omega_0). \tag{2.10}$$

The simplest array configuration is the **open sphere** composed of omnidirectional microphones suspended in free space. It is assumed that the microphones and associated cabling and mounting brackets are acoustically transparent, i.e., that they have no effect on the measured sound field. In this case, the mode strength is given by [89, 112]

$$b_l(k) = (-i)^l j_l(kr), \tag{2.11}$$

where $j_l(kr)$ is the spherical Bessel function of order l. This configuration is convenient for large array radii, where a rigid array would be impractical, and for scanning arrays. An example of an open spherical array, the *Sphere120 pro* by GFal, is shown in Fig. 2.3.

When processing the eigenbeams captured using the spherical array, it is necessary

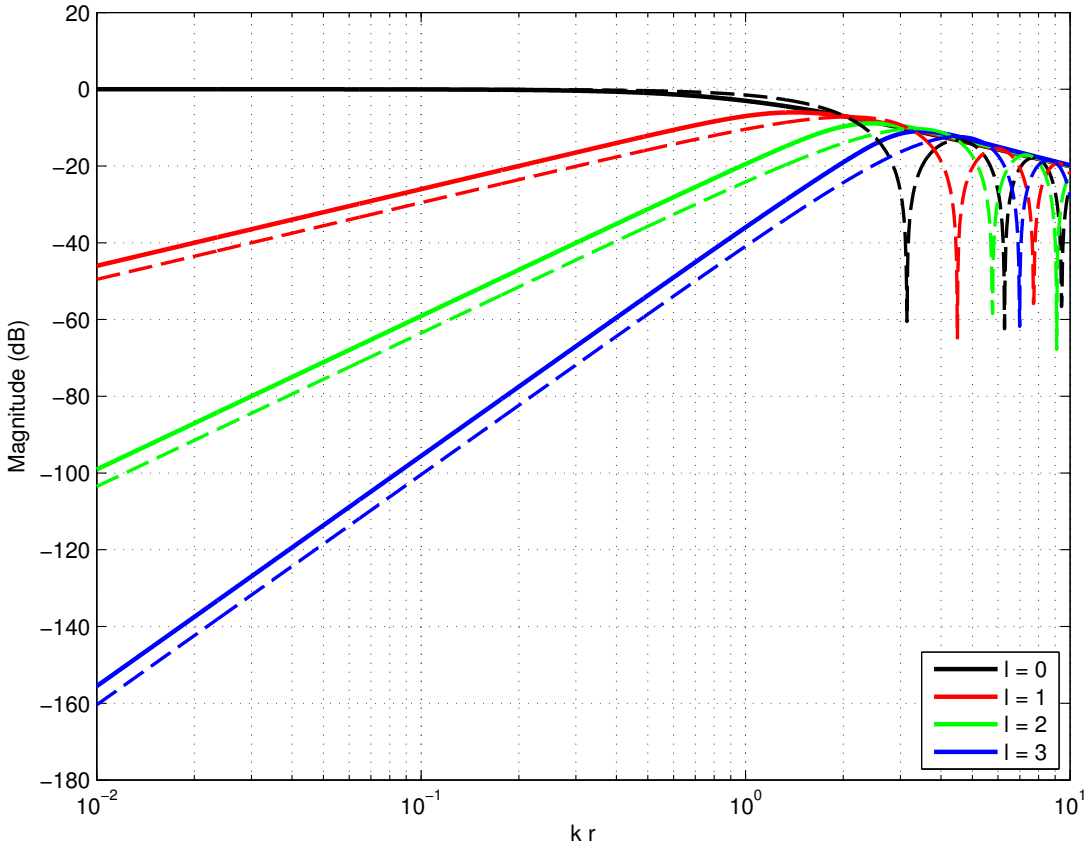


Figure 2.4: Magnitude of the mode strength $b_l(k)$ for orders $l \in \{0, 1, 2, 3\}$ as a function of kr. The solid lines denote a rigid sphere, and the dashed lines denote an open sphere.

to remove the dependence on the array properties by dividing the eigenbeams by $b_l(k)$, thereby removing the frequency-dependence of the eigenbeams. The open sphere mode strength is plotted in Figure 2.4 (dashed line); it can be seen that there are zeros at certain frequencies (for certain values of kr). As a result, the open array may suffer from poor robustness at these frequencies, where measurement noise will be significantly amplified. In addition, it can be seen that for $l > 0$, at low frequencies the mode strength is very small; as a result, high-order eigenbeams are generally not used at low frequencies [76].

The **rigid sphere** is a popular alternative to the open sphere. In this configuration, omnidirectional microphones are mounted on a rigid baffle, and the array is therefore no longer acoustically transparent: the sound waves are scattered by the sphere. An example of a rigid spherical array, the *Eigenmike* [79], is shown in Fig. 2.5. The mode



Figure 2.5: The em32 Eigenmike spherical microphone array. This rigid array of radius 4.2 cm is comprised of 32 omnidirectional microphones. (Photo credit: Emanuël Habets)

strength for a rigid sphere of radius r_a is given by [77, 89]

$$b_l(kr_\mathrm{a}, kr) = (-i)^l \left(j_l(kr) - \frac{j_l'(kr_\mathrm{a})}{h_l^{(1)\prime}(kr_\mathrm{a})} h_l^{(1)}(kr) \right), \tag{2.12}$$

where j_l' and $h_l^{(1)\prime}$ respectively denote the first derivatives of j_l and $h_l^{(1)}$ with respect to the argument, and $h_l^{(1)}$ is the spherical Hankel function of the first kind. The microphones are normally positioned on the surface of the rigid sphere (i.e., $r = r_\mathrm{a}$), therefore we define $b_l(k) \triangleq b_l(kr, kr)$. The second term in (2.12) compared to (2.11) accounts for the effect of scattering.

From the plot of the rigid sphere mode strength in Figure 2.4 (solid line), an advantage of the rigid sphere can be observed: it does not suffer from zeros in its mode strength, unlike the open sphere. In addition, the scattering effects of the rigid sphere are rigorously calculable and can be incorporated into the eigenbeam processing framework. For a detailed discussion of the scattering effects of the rigid sphere, the reader is referred to Chapter 3.

As the spherical microphone array available at the host institution is a rigid array (the *Eigenmike*), this configuration will be used for most of the work in this thesis. A number of other configurations have been proposed, but will not be discussed in this thesis. The mode strength expressions for the following configurations can be found in [93] and the references therein. The hemisphere [72] exploits the symmetry of the sound field by mounting the array on a rigid surface. The open dual-sphere [9], comprised of two spheres with different radii, and the open sphere with cardioid microphones [9] both overcome the problem of zeros in the open sphere mode strength, although cardioid microphones are not as readily available as omnidirectional microphones. Finally, in the free sampling configuration the microphones can be placed anywhere on the surface of a rigid sphere [71]; their positions are then optimized to robustly achieve an optimal approximation of a desired beampattern, or maximum directivity. The choice of array configuration is usually based on the intended application; for example, in a conference room where the microphone array is placed on a large table, the hemispherical configuration could be the most appropriate.

2.5 Beamforming

Once the sound field has been sampled and the eigenbeams have been computed, the eigenbeams can be combined to produce an enhanced output by applying aSHD beamformer. The output $Z(k)$ of an L^{th}-order SHD beamformer can be expressed as [90, eqn. 12]

$$Z(k) = \sum_{l=0}^{L} \sum_{m=-l}^{l} W_{lm}^{*}(k) P_{lm}(k), \tag{2.13}$$

where $W_{lm}(k)$ denotes the beamformer weights. The beamformer weights are chosen in order to achieve specific performance objectives.

The simplest beamformer is the **plane-wave decomposition beamformer** for which

the weights are given by [92]

$$W_{lm}^{*}(k) = \frac{Y_{lm}(\Omega_{\mathrm{u}})}{b_l(k)}, \tag{2.14}$$

where Ω_{u} is the beamformer look direction. As the array order L tends to infinity, the beamformer performs plane wave decomposition: the output tends towards a delta function in the direction of arrival (DOA) Ω_0 [89], i.e.,

$$\lim_{L\to\infty} Z(k) = \delta(\Theta), \tag{2.15}$$

where $\delta(\cdot)$ is the Dirac delta function and Θ is the angle between Ω_0 and Ω_{u}. The advantage of this beamformer is that it achieves maximum directivity [35, 95], i.e., the ratio of the output power in the look direction to the output power averaged over all directions [116] is maximized.

A commonly used beamformer in the spatial domain is the **delay-and-sum beamformer** (DSB), where it is assumed that the signals reaching each microphone in an array are identical with the exception of a time delay, and the beamformer output is formed by time-aligning and then summing the microphone signals [17, 116]. In the SHD, the so-called DSB is actually only mathematically equivalent to the spatial domain DSB in the case of an open sphere as $L \to \infty$ [91, eqn. 14]. Under these conditions, this beamformer achieves maximum white noise gain, i.e., the improvement in signal-to-noise ratio (SNR) between the array output and input for spatially white noise [116] is maximized [95]. The weights of the SHD DSB are given by [91, eqn. 16]

$$W_{lm}^{*}(k) = b_l^{*}(k) Y_{lm}(\Omega_{\mathrm{u}}). \tag{2.16}$$

A number of other more complex beamformers have also been proposed: some are fixed (like the two aforementioned beamformers), and apply a constraint to a specific look direction while optimizing the weights with respect to array performance measures (like

the directivity and white noise gain), whereas others are signal-dependent (e.g., [84,120]), and optimize the weights taking into account characteristics of the desired signal and noise. Rafaely provides a summary of some fixed beamforming methods in [92]. In Chapter 5, we propose a signal-dependent beamformer for noise reduction.

2.6 Associated literature

The main literature relevant to spherical microphone arrays has been referenced throughout this chapter, while the literature relevant to the specific problems addressed in this thesis (acoustic impulse response simulation, acoustic parameter estimation and signal enhancement) will be discussed in the relevant thesis chapters. The following section provides a brief overview of the fundamental publications in the field.

The literature relating to spherical microphone arrays is relatively sparse, and only begins in earnest at the turn of the century. Meyer & Elko, Gover, Ryan & Stinson, and Abhayapala & Ward were among the first to investigate spherical microphone arrays in 2002. Meyer & Elko presented an array based on a rigid sphere to be used for beamforming [77]. Gover, Ryan & Stinson used a spherical array to analyze acoustic impulse responses, reverberation times and the diffuseness of sound fields in rooms [43]. Abhayapala & Ward presented an open sphere array as an alternative to the *Soundfield* microphone [106] (a tetrahedral array composed of four microphones), capable of recording higher-order (second-order and above) sound fields, which they considered to be necessary for the accurate reproduction of a sound field [1].

The mathematical framework used for spherical array processing, based on spherical harmonic decomposition, was developed by Williams in *Fourier Acoustics* [119], where he gave a theoretical background on sound radiation with Fourier analysis in mind. In particular he rederived the equations that describe the scattering effect introduced by a rigid sphere, first formulated by Rayleigh in the 19$^{\text{th}}$ century [73]. Rafaely later presented a comprehensive theoretical analysis of spherical microphone arrays [90] and looked at

design issues such as sampling schemes, errors introduced by having a finite number of microphones, errors in microphone positioning, spatial aliasing, etc.

While the field of spherical array processing is relatively new, the spherical harmonics used are not: they were first introduced by Laplace in 1784, and are thus sometimes known as Laplace coefficients, despite the fact that the similar coefficients for two dimensions had been published by Legendre the previous year [97]. Since then they have been widely used in fields such as atomic physics, quantum chemistry, geodesy, magnetics, and computer graphics.

Although not the focus of this thesis, spherical microphone arrays can also be used for sound field recording and reproduction. *Ambisonics*, a series of surround sound acquisition and reproduction techniques, works with signals that are also based on a spherical harmonic decomposition of the sound field, although the terminology used is often different. Historically it has usually involved only zero- and first-order eigenbeams, referred to as *B-format* signals, although more recently higher-order systems have been investigated [80], providing increased spatial resolution. An introduction to Ambisonics is provided in [40].

Chapter 3

Acoustic impulse response simulation

In general, the evaluation of acoustic signal processing algorithms, such as direction of arrival (DOA) estimation (see Chapter 4) and speech enhancement (see Chapter 5) algorithms, makes use of simulated acoustic transfer functions (ATFs). By using simulated ATFs it is possible to comprehensively evaluate an algorithm under many acoustic conditions (e.g., reverberation time, room dimensions and source-array distance). Allen & Berkley's image method [6] is a widely used approach to simulate ATFs between an omnidirectional sound source and one or more microphones in a reverberant environment. In the last few decades, several extensions have been proposed [70, 85].

In recent years the use of spherical microphone arrays has become prevalent. These arrays are commonly of one of two types: the open array, where microphones are suspended in free space on an 'open' sphere, and the rigid array, where microphones are mounted on a rigid baffle. The rigid sphere is often preferred as it improves the numerical stability of many processing algorithms [89] and its scattering effects are rigorously calculable [77].

Currently, many works relating to spherical array processing consider only free-field responses, however, when a rigid array is used, the rigid baffle causes scattering of the

---

Portions of this work were first published in the *Journal of the Acoustical Society of America* [62] in 2012. © 2012 Acoustical Society of America.

sound waves incident upon the array that the image method does not consider. This scattering has an effect on the ATFs, especially at high frequencies and/or for microphones situated on the occluded side of the array. Furthermore the reverberation due to room boundaries such as walls, ceiling and floor must also be considered, particularly in small rooms.

While measured transfer functions include both these effects, they are both time-consuming and expensive to acquire. A method for simulating ATFs in a reverberant room while accounting for scattering is therefore essential, allowing for fast, comprehensive and repeatable testing. In this chapter, we propose such a method that combines a model of the scattering in the spherical harmonic domain (SHD) with a version of the image method that accounts for reverberation in a computationally efficient way.

The simulated ATFs include the direct path, reflections due to room reverberation, scattering of the direct path and scattering of the reverberant reflections. Reflections of the scattered sound and multiple interactions between the room boundaries and the sphere are excluded as they do not contribute significantly to the sound field, provided the distances between the room boundaries and the sphere are several times the sphere's radius [44], which is easily achieved in the case of a small scatterer [16]. Furthermore, we assume an empty rectangular shoebox room (with the exception of the rigid sphere) and specular reflections, as was assumed in the conventional image method [6]. Finally, the scattering model used assumes a perfectly rigid baffle, without absorption.

In this chapter, we first briefly summarize Allen & Berkley's image method and then present our proposed method in the SHD. We then discuss some implementation aspects, namely the truncation of an infinite sum in the ATF expression and the reduction of the method's computational complexity, and then provide a pseudocode description of the method. An open-source software implementation is available online [54]. Finally, we show some example uses of the method and, where possible, compare the simulated results obtained with theoretical models. Earlier versions of this work were previously published in [61, 62].

3.1 Allen & Berkley's image method

The source-image or image method [6] is one of the most commonly used room acoustics simulation methods in the acoustic signal processing community. The principle of the method is to model an ATF as the sum of a direct path component and a number of discrete reflections, each of these components being represented in the ATF by a free-space Green's function. In this section, we review the free-space Green's function and the image method.

3.1.1 Green's function

For a source at a position $\mathbf{r}_s$ and a receiver at a position $\mathbf{r}$, the free-space Green's function, a solution to the inhomogeneous Helmholtz equation applying the Sommerfeld radiation condition, is given by[1]

$$G(\mathbf{r}|\mathbf{r}_s, k) = \frac{e^{+ik\|\mathbf{r}-\mathbf{r}_s\|}}{4\pi\|\mathbf{r}-\mathbf{r}_s\|}, \tag{3.1}$$

where $\|\cdot\|$ denotes the ℓ-2 norm and the wavenumber k is related to frequency f (in Hz), angular frequency ω (in $\mathrm{rad}\cdot\mathrm{s}^{-1}$) and the speed of sound c (in $\mathrm{m}\cdot\mathrm{s}^{-1}$) via the relationship $k = \omega/c = 2\pi f/c$.

In the time-domain, the Green's function is given by

$$g(\mathbf{r}|\mathbf{r}_s, t) = \frac{\delta(t - \frac{\|\mathbf{r}-\mathbf{r}_s\|}{c})}{4\pi\|\mathbf{r}-\mathbf{r}_s\|}, \tag{3.2}$$

where δ is the Dirac delta function and t is time. This corresponds to a pure impulse at time $t = \frac{\|\mathbf{r}-\mathbf{r}_s\|}{c}$, i.e. the propagation time from $\mathbf{r}_s$ to $\mathbf{r}$.

[1] This expression assumes the sign convention commonly used in physics/acoustics, whereby the temporal Fourier transform is defined as $F(\omega) = \int_{-\infty}^{\infty} f(t)e^{+i\omega t}\mathrm{d}t$ in order to eliminate the $e^{-i\omega t}$ term in the time-harmonic solution to the wave equation, as in Morse & Ingard [81] and Williams [119]. The formulae in this thesis are the complex conjugates of those found in other publications which use the opposite sign convention.

3.1.2 Image method

Consider a rectangular room with length L_x, width L_y and height L_z. The reflection coefficients of the four walls, floor and ceiling are β_{x_1}, β_{x_2}, β_{y_1}, β_{y_2}, β_{z_1} and β_{z_2}, where the v_1 coefficients ($v \in \{x, y, z\}$) correspond to the boundaries at $v = 0$ and the v_2 coefficients correspond to the boundaries at $v = L_v$.

If the sound source is located at $\mathbf{r}_s = (x_s, y_s, z_s)$ and the receiver is located at $\mathbf{r} = (x, y, z)$, the images obtained using the walls at $x = 0$, $y = 0$ and $z = 0$ can be expressed as a vector $\mathbf{R_p}$:

$$\mathbf{R_p} = [x_s - x + 2p_x x, y_s - y + 2p_y y, z_s - z + 2p_z z], \tag{3.3}$$

where each of the elements in $\mathbf{p} = (p_x, p_y, p_z)$ can take values 0 or 1, thus resulting in eight combinations that form a set $\mathcal{P}$. To consider all reflections we also define a vector $\mathbf{R_m}$ which we add to $\mathbf{R_p}$:

$$\mathbf{R_m} = [2m_x L_x, 2m_y L_y, 2m_z L_z], \tag{3.4}$$

where each of the elements in $\mathbf{m} = (m_x, m_y, m_z)$ can take values between $-N_m$ and N_m, and N_m is used to limit computational complexity and circular convolution errors, thus resulting in a set $\mathcal{M}$ of $(2N_m + 1)^3$ combinations. The image positions in the x and y dimensions are illustrated in Fig. 3.1.

The distance between an image and the receiver is given by $\left\|\mathbf{R_p} + \mathbf{R_m}\right\|$. Using (3.1), the ATF H is then given by

$$H(\mathbf{r}|\mathbf{r}_s, k) = \sum_{\mathbf{p}\in\mathcal{P}} \sum_{\mathbf{m}\in\mathcal{M}} \beta_{x_1}^{|m_x+p_x|} \beta_{x_2}^{|m_x|} \beta_{y_1}^{|m_y+p_y|} \beta_{y_2}^{|m_y|} \beta_{z_1}^{|m_z+p_z|} \beta_{z_2}^{|m_z|} \frac{e^{+ik\left\|\mathbf{R_p}+\mathbf{R_m}\right\|}}{4\pi\left\|\mathbf{R_p} + \mathbf{R_m}\right\|}. \tag{3.5}$$

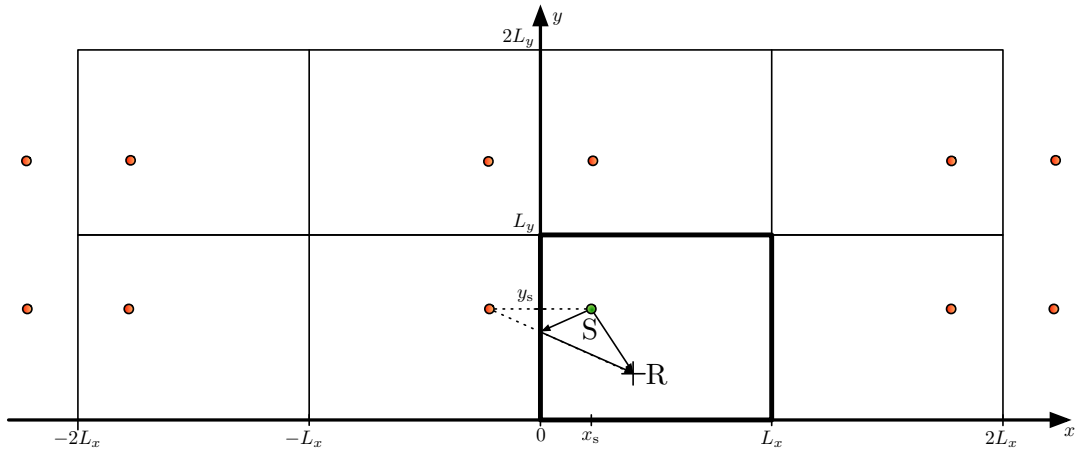


Figure 3.1: A slice through the image space showing the positions of the images in the x and y dimensions, with a source S and receiver R. The full image space has three dimensions (x, y and z). An example of a reflected path (first-order reflection about the x-axis) is also shown.

Using (3.2), we obtain the acoustic impulse response (AIR)

$$h(\mathbf{r}|\mathbf{r}_\mathrm{s}, t) = \sum_{\mathbf{p}\in\mathcal{P}} \sum_{\mathbf{m}\in\mathcal{M}} \beta_{x_1}^{|m_x+p_x|} \beta_{x_2}^{|m_x|} \beta_{y_1}^{|m_y+p_y|} \beta_{y_2}^{|m_y|} \beta_{z_1}^{|m_z+p_z|} \beta_{z_2}^{|m_z|} \frac{\delta\left(t - \frac{\left\|\mathbf{R}_\mathbf{p}+\mathbf{R}_\mathbf{m}\right\|}{c}\right)}{4\pi\left\|\mathbf{R}_\mathbf{p}+\mathbf{R}_\mathbf{m}\right\|}. \tag{3.6}$$

3.2 Proposed method in the spherical harmonic domain

There exists a compact analytical expression for the scattering due to the rigid sphere in the SHD, therefore we first express the free-space Green's function in this domain, and then use this to form an expression for the ATF including scattering.

3.2.1 Green's function

We define position vectors in spherical coordinates relative to the centre of our array. Letting r be the array radius and Ω an inclination-azimuth pair, the microphone position vector is defined as $\tilde{\mathbf{r}} \triangleq (r, \Omega)$ (where in this chapter $\tilde{\cdot}$ indicates a vector in spherical coordinates). Similarly, the source position vector is given by $\tilde{\mathbf{r}}_\mathrm{s} \triangleq (r_\mathrm{s}, \Omega_\mathrm{s})$. It is hereafter assumed that where the addition, ℓ-2 norm or scalar product operations are applied to spherical polar vectors, they have previously been converted to Cartesian coordinates.

The free-space Green's function (3.1) can be expressed in the SHD using the following spherical harmonic decomposition [119]:

$$\begin{aligned} G(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_\mathrm{s}, k) =& \frac{e^{+ik\|\tilde{\mathbf{r}}-\tilde{\mathbf{r}}_\mathrm{s}\|}}{4\pi\|\tilde{\mathbf{r}}-\tilde{\mathbf{r}}_\mathrm{s}\|} \\ =& ik\sum_{l=0}^{\infty}\sum_{m=-l}^{l} j_l(kr)h_l^{(1)}(kr_\mathrm{s})Y_{lm}^*(\Omega_\mathrm{s})Y_{lm}(\Omega) \\ =& ik\sum_{l=0}^{\infty} j_l(kr)h_l^{(1)}(kr_\mathrm{s})\sum_{m=-l}^{l} Y_{lm}^*(\Omega_\mathrm{s})Y_{lm}(\Omega), \end{aligned} \tag{3.7}$$

where Y_{lm} is the spherical harmonic function of order l and degree m, j_l is the spherical Bessel function of order l and $h_l^{(1)}$ is the spherical Hankel function of the first kind and of order l. This decomposition is also known as a spherical Fourier series expansion or spherical harmonics expansion of the Green's function.

According to the spherical harmonic addition theorem [119],

$$\sum_{m=-l}^{l} Y_{lm}^*(\Omega_\mathrm{s})Y_{lm}(\Omega) = \frac{2l+1}{4\pi}\mathcal{P}_l(\cos\Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{r}}_\mathrm{s}}), \tag{3.8}$$

where $\mathcal{P}_l$ is the Legendre polynomial of order l and $\Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{r}}_\mathrm{s}}$ is the angle between $\tilde{\mathbf{r}}$ and $\tilde{\mathbf{r}}_\mathrm{s}$. Using this theorem, which in many cases reduces the complexity of the implementation, we can simplify the Green's function in (3.7) to

$$G(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_\mathrm{s}, k) = \frac{ik}{4\pi}\sum_{l=0}^{\infty} j_l(kr)h_l^{(1)}(kr_\mathrm{s})(2l+1)\mathcal{P}_l(\cos\Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{r}}_\mathrm{s}}). \tag{3.9}$$

The cosine of the angle $\Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{r}}_\mathrm{s}}$ is obtained as the dot product of the two normalized vectors $\hat{\mathbf{r}}_\mathrm{s} = \tilde{\mathbf{r}}_\mathrm{s}/r_\mathrm{s}$ and $\hat{\mathbf{r}} = \tilde{\mathbf{r}}/r$:

$$\cos\Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{r}}_\mathrm{s}} = \hat{\mathbf{r}}\cdot\hat{\mathbf{r}}_\mathrm{s}. \tag{3.10}$$

3.2.2 Neumann Green's function

The free-space Green's function describes the propagation of sound in free space only. However, when a rigid sphere is present, a boundary condition must hold: the radial

velocity must vanish on the surface of the sphere. The function $G_N(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_s, k)$ satisfying this boundary condition is called the *Neumann Green's* function, and describes the sound propagation between a point $\tilde{\mathbf{r}}_s$ and a point $\tilde{\mathbf{r}}$ on the rigid sphere [119]:

$$\begin{aligned} G_N(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_s, k) &= G(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_s, k) - \frac{ik}{4\pi}\sum_{l=0}^{\infty}\frac{j_l'(kr)}{h_l^{(1)\prime}(kr)}h_l^{(1)}(kr)h_l^{(1)}(kr_s)(2l+1)\mathcal{P}_l(\cos\Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{r}}_s}) \\ &= \frac{k}{4\pi}\sum_{l=0}^{\infty}(-i)^{-(l+1)}b_l(k)h_l^{(1)}(kr_s)(2l+1)\mathcal{P}_l(\cos\Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{r}}_s}), \end{aligned} \tag{3.11}$$

where $(\cdot)'$ denotes the first derivative and the term

$$b_l(k) = (-i)^l\left(j_l(kr) - \frac{j_l'(kr)}{h_l^{(1)\prime}(kr)}h_l^{(1)}(kr)\right) \tag{3.12}$$

is often called the (farfield) *mode strength*. For the open sphere, $b_l(k) = (-i)^l j_l(kr)$ yields the free-space Green's function. The Wronskian relation for the spherical Bessel and Hankel functions [119]

$$j_l(x)h_l^{(1)\prime}(x) - j_l'(x)h_l^{(1)}(x) = \frac{i}{x^2} \tag{3.13}$$

allows us to simplify (3.12) to

$$b_l(k) = \frac{(-i)^{l-1}}{h_l^{(1)\prime}(kr)(kr)^2}. \tag{3.14}$$

3.2.3 Scattering model

The proposed rigid sphere scattering model[2] has a long history in the literature; it was first developed by Clebsch and Rayleigh in 1871-72 [73]. It is presented in a number of acoustics texts [81, 101, 119], and is used in many theoretical analyses for spherical

[2]Some texts [33] refer to the scattering effect as diffraction, although Morse & Ingard note that *"When the scattering object is large compared with the wavelength of the scattered sound, we usually say the sound is reflected and diffracted, rather than scattered"* [81], therefore in the case of spherical microphone arrays (particularly rigid ones which tend to be relatively small for practical reasons), scattering is possibly the more appropriate term.

microphone arrays [78, 90].

3.2.3.1 Theoretical behaviour

The behaviour of the scattering model is illustrated in Fig. 3.2, which plots the magnitude of the response between a source and a receiver on a rigid sphere of radius 5 cm for a source-array distance of 1 m, as a function of frequency and DOA. The response was normalized using the free-field/open sphere response, therefore the plot shows only the effect due to scattering. Due to rotational symmetry, we only looked at the one-dimensional DOA, instead of looking at both azimuth and inclination, and limited the DOA to the 0–180° range.

When the source is located on the same side of the sphere as the receiver (i.e. the direction of arrival is 0°), the rigid sphere response is greater than the open sphere response due to constructive scattering, tending towards a 6 dB magnitude gain compared to the open sphere at infinite frequency. The response on the back side of the rigid sphere is generally lower than in the open sphere case and lower than on the front side, as one would intuitively expect due to it being occluded. However at the very back of the sphere (i.e. the DOA is 180°) we observe a narrow *bright spot*: the waves propagating around the sphere all arrive in phase at the 180° point and as a result sum constructively.

A polar plot of the magnitude response (Fig. 3.3) illustrates both the near-doubling of the response on the front side of the sphere, and the bright spot on the back side of the sphere, which narrows as frequency increases. It should be noted that although the above results are for a fixed sphere radius, as the scattering model is a function of kr, the effects of a change in radius are the same as a change in frequency; indeed the relevant factor is the radius of the sphere relative to the wavelength.

These substantial differences between the open and rigid sphere responses confirm the need for a simulation method which accounts for scattering, even for sphere radii as small as 5 cm.

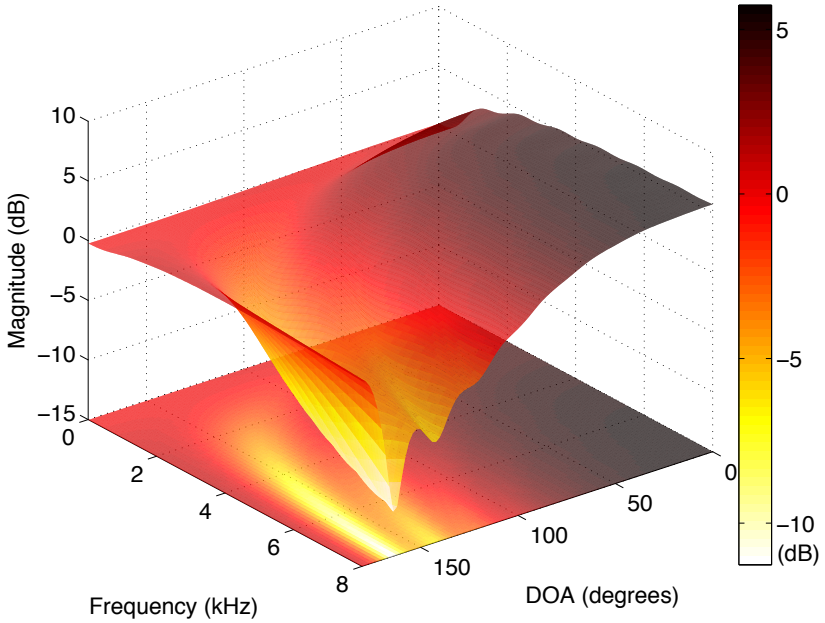


Figure 3.2: Magnitude of the response between a source and a receiver placed on a rigid sphere of radius 5 cm at a distance of 1 m, as a function of frequency and DOA. The response was normalized with respect to the free-field response.

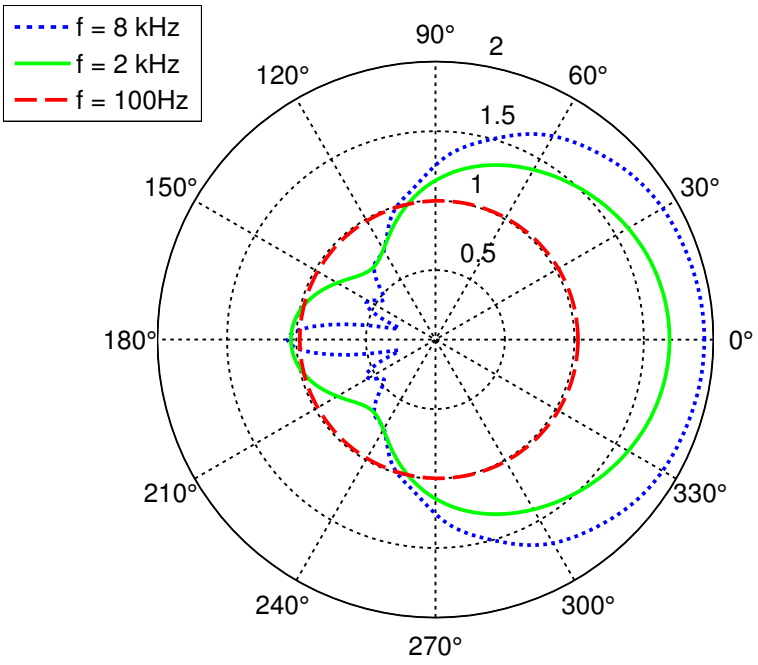


Figure 3.3: Polar plot of the magnitude of the response between a source and a receiver placed on a rigid sphere of radius 5 cm, at a distance of 1 m, for various frequencies.

3.2.3.2 Experimental validation

In addition to being widely used in theory, this model has also been experimentally validated by Duda & Martens [33] using a single microphone inserted in a hole drilled through a 10.9 cm radius bowling ball placed in an anechoic chamber. This is a reasonable approximation to a spherical microphone array; indeed a bowling ball was used by Li & Duraiswami to construct a hemispherical microphone array [72].

Duda & Martens's experimental results broadly agree with the theoretical model. In our case we are most interested in the results in their Fig. 12 a) where the source-array distance is largest (20 times the array radius), as in typical spherical array usage scenarios the source is unlikely to be much closer to the array. The only notable difference between the theoretical and experimental results in this case is for a direction of arrival of 180°, where the high frequency response is found to be lower than expected. The authors suggest this is due to small alignment errors, which would indeed have an effect given the narrowness of the bright spot in the model (see Fig. 3.3 for $f = 8$ kHz). Given these results, we conclude that the use of this scattering model is sufficiently accurate for simulating a small rigid array, such as the *Eigenmike* [79].

3.2.4 Proposed method

We now present our proposed method, incorporating the spherical harmonic decomposition presented in Section 3.2.1 and the scattering model introduced in Section 3.2.2.

Due to the differences between the SHD Neumann Green's function in (3.11) and the spatial domain Green's function in (3.1), as well as the directionality of the array's response, two changes must be made to the image position vectors $\mathbf{R_p}$ and $\mathbf{R_m}$ in our proposed method. Firstly, to compute the spherical harmonic decomposition in the Neumann Green's function, we require the distance between each image and the *centre* of the array [r_s in (3.11)]; this is accomplished by computing the image position vectors using the position of the centre of the array rather than the position of the receiver. Secondly,

to compute the spherical harmonic decomposition we require the angle between each image and the receiver with respect to the centre of the array [$\Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{r}}_s}$ in (3.11)]. In Allen & Berkley's image method, the direction of the vector $\mathbf{R_p} + \mathbf{R_m}$ is not always the same: in some cases it points from the receiver to the image and in others it points from the image to the receiver. This is not an issue for the image method as only the norm of this vector is used. As we also require the angle of the images in our proposed method, we modify the definition of $\mathbf{R_p}$ such that the vector $\mathbf{R_p} + \mathbf{R_m}$ always points from the centre of the array to the image.

We now incorporate these two changes into the definition of the image vectors $\mathbf{R_p}$ and $\mathbf{R_m}$. If the sound source is located at $\mathbf{r}_s = (x_s, y_s, z_s)$ and the centre of the sphere is located at $\mathbf{r}_a = (x_a, y_a, z_a)$, the images obtained using the walls at $x = 0$, $y = 0$ and $z = 0$ are expressed as a vector $\mathbf{R_p}$:

$$\mathbf{R_p} = [x_s - 2p_x x_s - x_a, y_s - 2p_y y_s - y_a, z_s - 2p_z z_s - z_a]. \tag{3.15}$$

For brevity we define $\mathbf{R}_{\mathbf{p},\mathbf{m}} \triangleq \mathbf{R_p} + \mathbf{R_m}$, allowing us to express the distance between an image and the centre of the sphere as $\left\|\mathbf{R}_{\mathbf{p},\mathbf{m}}\right\|$ and the angle between the image and the receiver as $\Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{R}}_{\mathbf{p},\mathbf{m}}}$, computed in the same way as (3.10), where $\tilde{\mathbf{R}}_{\mathbf{p},\mathbf{m}}$ denotes the image positions in spherical coordinates. The image positions in the x dimension are illustrated in Fig. 3.4. Finally, the ATF $H(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_s, k)$ is the weighted sum of the individual responses $G_N(\tilde{\mathbf{r}}|\tilde{\mathbf{R}}_{\mathbf{p},\mathbf{m}}, k)$ for each of the images[3]

$$H(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_s, k) = \sum_{\mathbf{p}\in\mathcal{P}} \sum_{\mathbf{m}\in\mathcal{M}} \beta_{x_1}^{|m_x - p_x|} \beta_{x_2}^{|m_x|} \beta_{y_1}^{|m_y - p_y|} \beta_{y_2}^{|m_y|} \beta_{z_1}^{|m_z - p_z|} \beta_{z_2}^{|m_z|} G_N(\tilde{\mathbf{r}}|\tilde{\mathbf{R}}_{\mathbf{p},\mathbf{m}}, k). \tag{3.16}$$

Since we are working in the wavenumber domain, we can allow for frequency dependent boundary reflection coefficients in (3.16), if desired. The reflection coefficients would then be written as $\beta_{x_1}(k)$, $\beta_{x_2}(k)$ and so on. Chen & Maher [21] provide some

[3]The sign in the powers of β is different from that in Allen & Berkley's conventional image method, due to the change in the definition of $\mathbf{R_p}$ that is required for our proposed method.

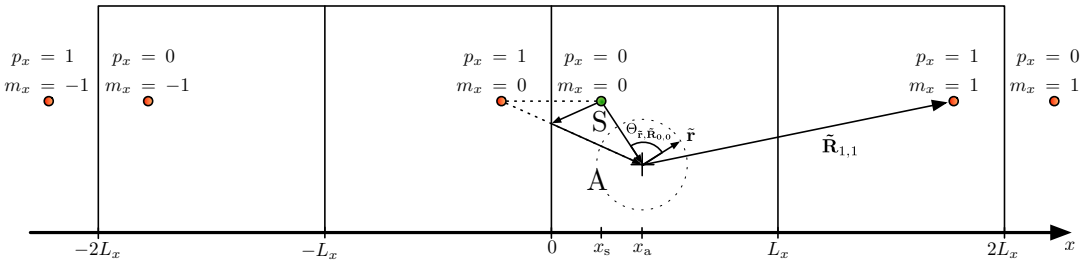


Figure 3.4: A slice through the image space showing the positions of the images in the x dimension, with a source S and array A. The full image space has three dimensions (x, y and z). An example of a reflected path is shown for the image with $p_x = 1$ and $m_x = 0$.

measured reflection coefficients for a wall, window, floor and ceiling.

3.3 Implementation

3.3.1 Truncation error

To compute the expression for the ATF in (3.16), the sum over an infinite number of orders l in the Neumann Green's function G_N must be approximated by a sum $\hat{G}_N$ over a finite order L. Choosing L too small will result in a large approximation error, while choosing L too large will result in too high a computational complexity. We now investigate the approximation error in order to provide some guidelines for the choice of the order L. The results for an open sphere are provided for reference, and were computed by using a truncated spherical harmonic decomposition of the Green's function $\hat{G}$ instead of a Neumann Green's function.

For an open sphere, the error can be determined exactly because the Green's function is a decomposition of the closed-form expression in (3.1). For a rigid sphere, however, no closed-form expression exists since the scattering term can be expressed only in the SHD. We therefore estimated the error by comparing the truncated Neumann Green's function $\hat{G}_N$ to a high-order Neumann Green's function. Based on simulations performed with an open sphere, where a true reference is available, we can safely assume that the error involved in using a high-order Neumann Green's function as a reference as opposed to the

untruncated Neumann Green's function is small. In practice, we cannot choose very large values of L because of numerical difficulties involved in multiplying high order spherical Bessel and Hankel functions. For typical sphere radii and source-array distances, this allows us to reach L values of up to about 100 using our MATLAB implementation [54].

We evaluated the truncated (Neumann) Green's function at $K = 1024$ discrete values of k (denoted by $\dot{k}$), forming a set $\mathcal{K}$ corresponding to frequencies in the range 100 Hz - 8 kHz[4], and then calculated the normalized root-mean-square magnitude error ϵ_{m} and the root-mean-square phase error ϵ_{p}, i.e.,

$$\epsilon_{\mathrm{m}}(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_{\mathrm{s}}, L) = \sqrt{\frac{1}{K}\sum_{\dot{k}\in\mathcal{K}} \frac{\left(\left|G_{\mathrm{N}}(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_{\mathrm{s}}, \dot{k})\right| - \left|\hat{G}_{\mathrm{N}}(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_{\mathrm{s}}, \dot{k}, L)\right|\right)^2}{\left|G_{\mathrm{N}}(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_{\mathrm{s}}, \dot{k})\right|^2}}, \tag{3.17}$$

$$\epsilon_{\mathrm{p}}(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_{\mathrm{s}}, L) = \sqrt{\frac{1}{K}\sum_{\dot{k}\in\mathcal{K}} \left(\angle G_{\mathrm{N}}(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_{\mathrm{s}}, \dot{k}) - \angle \hat{G}_{\mathrm{N}}(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_{\mathrm{s}}, \dot{k}, L)\right)^2}. \tag{3.18}$$

We averaged the magnitude and phase errors over 32 quasi-equidistant receivers and 50 random source positions at a fixed distance from the centre of the array.

The resulting average errors are given in Fig. 3.5, for both the open and rigid sphere cases. Three different sphere radii were used: $r = 4.2$ cm (the radius of the *Eigenmike* [76]), $r = 10$ cm and $r = 15$ cm. A source-array distance of 1 m was used; results for 1–5 m are omitted as they are essentially identical. It can be seen that beyond a certain threshold, increases in L give only a very small reduction in error; this is due to the fast convergence of the spherical harmonic decomposition [45]. From Fig. 3.5, we can see that a sensible rule of thumb for choosing L is $L > \lceil 1.1\, k_{\max} r \rceil$ where $k_{\max}$ is the largest wavenumber of interest.

[4] Very low frequencies are omitted due to the fact that the spherical Hankel function $h_l(x)$ has a singularity around $x = 0$.

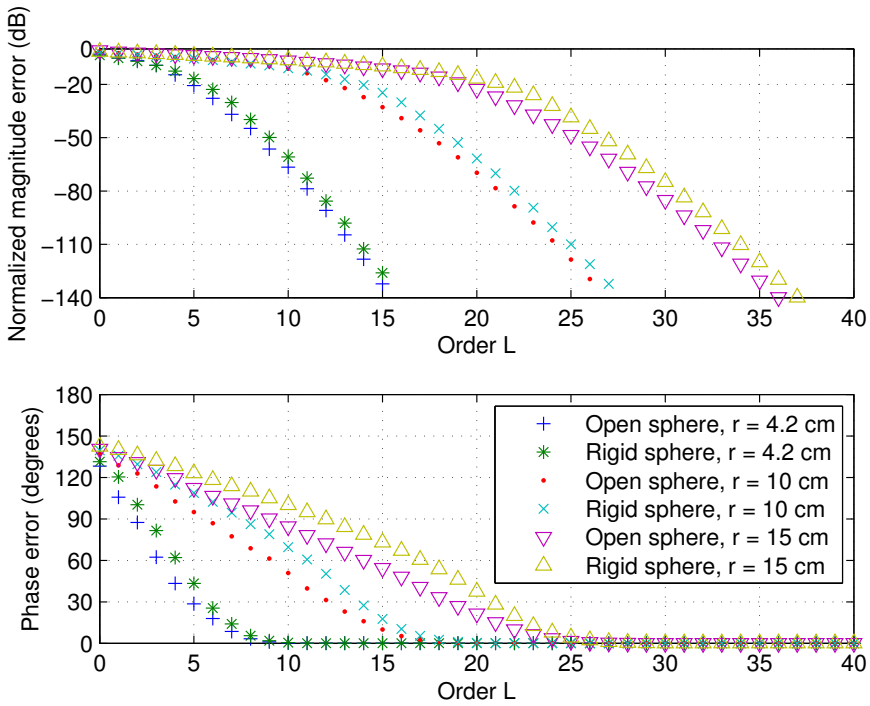


Figure 3.5: Magnitude and phase errors involved in the truncation of the spherical harmonic decomposition in the Green's function (open sphere) and the Neumann Green's function (rigid sphere). The errors reduce rapidly beyond $L = k_{\max} r$, where here $k_{\max} = 147\ \mathrm{m}^{-1}$.

3.3.2 Computational complexity

As the ATFs are made up of a sum over all orders l which includes spherical Hankel functions h_l and Legendre polynomials $\mathcal{P}_l$, we can make use of recursion relations over l to reduce the computational complexity of these functions. For the spherical Hankel function, we make use of the following relation [2]

$$h_m(x) = \frac{2m-1}{x} h_{m-1}(x) - h_{m-2}(x), \ m \geq 2 \tag{3.19}$$

where

$$h_0(x) = \frac{e^{ix}}{ix}, \ h_1(x) = \frac{e^{ix}}{ix^2} - \frac{e^{ix}}{x}. \tag{3.20}$$

For the Legendre polynomial we use a similar recursion relation [2], known as Bonnet's recursion formula

$$\mathcal{P}_m(x) = \frac{2m-1}{m} x \mathcal{P}_{m-1}(x) - \frac{m-1}{m} \mathcal{P}_{m-2}(x), \; m \geq 2 \tag{3.21}$$

where $\mathcal{P}_0(x) = 1$ and $\mathcal{P}_1(x) = x$.

While replacing the exponential in (3.1) with a spherical harmonic decomposition does lead to an increase in computational complexity when computing the ATF for a single receiver (which is unavoidable in the rigid sphere case), this can have an advantage when simulating many receiver positions. For the conventional image method, we must compute the image positions and resulting response separately for each individual receiver. However, in the proposed method the image positions are all computed with respect to the *centre* of our array, and therefore only once for all of the microphones in the array.

An alternative to (3.16) is obtained by changing the order of the summations in the ATF and computing the sum over all images only once, instead of once per receiver, i.e.,

$$\begin{aligned} H(\tilde{\mathbf{r}}|\tilde{\mathbf{r}}_s, k) = k \sum_{l=0}^{\infty} (-i)^{-(l+1)} \sum_{m=-l}^{l} Y_{lm}(\Omega) \\ \cdot \sum_{\mathbf{p}\in\mathcal{P}} \sum_{\mathbf{m}\in\mathcal{M}} \beta_{x_1}^{|m_x-p_x|} \beta_{x_2}^{|m_x|} \beta_{y_1}^{|m_y-p_y|} \beta_{y_2}^{|m_y|} \beta_{z_1}^{|m_z-p_z|} \beta_{z_2}^{|m_z|} b_l(k) h_l^{(1)}(k \left\| \mathbf{R}_{\mathbf{p},\mathbf{m}} \right\|) Y_{lm}^*(\angle \mathbf{R}_{\mathbf{p},\mathbf{m}}). \end{aligned} \tag{3.22}$$

The expression in (3.22) requires $O\left((N+Q)(L+1)^2\right)$ operations per discrete frequency, where L is the maximum spherical harmonic order, N is the number of images and Q is the number of microphones, while the approach in (3.16) requires $O\left(NQ(L+1)\right)$ operations per discrete frequency. Since the number of images N is typically very large, $(N+Q)(L+1)^2 \approx N(L+1)^2$. Assuming the operations in the two approaches are of similar complexity, it is therefore more efficient to use the expression in (3.16) for $Q < L+1$ and the expression in (3.22) for $Q > L+1$. Consequently the least computationally

complex approach depends on the number of microphones Q and array radius r. In the remainder of this chapter we use the expression in (3.16); this is particularly appropriate in the applications in Section 3.4.2 where $Q = 2$ and in Section 3.4.3 where $Q = 1$.

3.3.3 Algorithm summary

A summary of the proposed method is presented in the form of pseudocode in Fig. 3.6. The variable $nsample$ denotes the number of samples in the AIR and N_o the maximum reflection order.

The number of computations has been reduced by processing only half of the frequency spectrum because we know the AIR is real and the corresponding ATF is conjugate symmetric. The pseudocode necessary to compute the Hankel functions and Legendre polynomials is omitted here, since their computation is straightforward using recursion relations (3.19) and (3.21).

SMIRgen, a MATLAB/C++ implementation of the method in the form of a MEX-function is presented in Appendix A and is available online [54].

3.4 Examples and applications

In this section we give a number of examples that make use of the proposed method. Wherever possible we compared the simulated results to theoretical results obtained using approximate models. These examples are given to illustrate and partially validate the proposed method.

3.4.1 Diffuse sound field energy

In statistical room acoustics (SRA), reverberant sound fields are modelled as diffuse sound fields, allowing for a statistical analysis of reverberation instead of computing each of the individual reflections. In this subsection, we compare a theoretical prediction of

$\mathcal{P} = \{0,1\}^3$
$\mathcal{M} = \{-N_m, \cdots, 0, \cdots, N_m\}^3$
$\mathcal{A} = \mathcal{P} \times \mathcal{M}$

for $(\mathbf{p},\mathbf{m}) \in \mathcal{A}$ **do**
  if $|2m_x - p_x| + |2m_y - p_y| + |2m_z - p_z| \leq N_o$ **then**
    $\mathbf{R}_{\mathbf{p},\mathbf{m}} = \begin{bmatrix} x_\mathrm{s} - 2p_x x_\mathrm{s} - x_\mathrm{a} + 2m_x L_x \\ y_\mathrm{s} - 2p_y y_\mathrm{s} - y_\mathrm{a} + 2m_y L_y \\ z_\mathrm{s} - 2p_z z_\mathrm{s} - z_\mathrm{a} + 2m_z L_z \end{bmatrix}$
    $\beta(\mathbf{p},\mathbf{m}) = \beta_{x_1}^{|m_x - p_x|} \beta_{x_2}^{|m_x|} \beta_{y_1}^{|m_y - p_y|} \beta_{y_2}^{|m_y|} \beta_{z_1}^{|m_z - p_z|} \beta_{z_2}^{|m_z|}$
  else
    $\mathcal{A} = \mathcal{A} \setminus \{(\mathbf{p},\mathbf{m})\}$
  end if
end for

for $k = 1 \rightarrow nsample/2 + 1$ **do**
  for $l = 0 \rightarrow L$ **do**
    if $sphType =$'rigid' **then**
      $\Delta(k,l) = \frac{(-i)^{l-1}}{h_l^{(1)'}(kr)(kr)^2}$
    else
      $\Delta(k,l) = (-i)^l j_l(kr)$
    end if
  end for
  $\Gamma(k,l) = \frac{(-i)^{-(l+1)}k}{4\pi} \cdot \Delta(k,l)$
end for

for $(\mathbf{p},\mathbf{m}) \in \mathcal{A}$ **do**
  if $||\mathbf{R}_{\mathbf{p},\mathbf{m}}|| + r < c \cdot nsample/fs$ **then**
    for $ang = 1 \rightarrow Q$ **do**
      $\Theta = \hat{\mathbf{R}}_{\mathbf{p},\mathbf{m}} \cdot \hat{\mathbf{r}}(ang)$
      $\Psi = P_l(\Theta)$
      for $l = 0 \rightarrow L$ **do**
        $\Upsilon(ang,l) = \Psi \cdot (2l+1)$
      end for
    end for
    for $k = 1 \rightarrow nsample/2 + 1$ **do**
      for $l = 0 \rightarrow L$ **do**
        $\Lambda(k,l) = h_l(k||\mathbf{R}_{\mathbf{p},\mathbf{m}}||) \cdot \Gamma(k,l)$
      end for
    end for
    for $ang = 1 \rightarrow Q$ **do**
      for $k = 1 \rightarrow nsample/2 + 1$ **do**
        for $l = 0 \rightarrow L$ **do**
          $H(\mathbf{p},\mathbf{m},ang,k,l) = \beta(\mathbf{p},\mathbf{m}) \cdot \Upsilon(ang,l) \cdot \Lambda(k,l)$
        end for
        $H(\mathbf{p},\mathbf{m},ang,k) = \sum_l H(\mathbf{p},\mathbf{m},ang,k,l)$
      end for
    end for
  end if
end for
$H(ang,k) = \sum_{(\mathbf{p},\mathbf{m})\in\mathcal{A}} H(\mathbf{p},\mathbf{m},ang,k)$
$h(ang,n) = \mathrm{IFFT}_\mathrm{R}\{H(ang,k)\}$

Figure 3.6: Pseudocode for the proposed method.

sound energy on the surface of a rigid sphere, based on a diffuse model of reverberation, to simulated results obtained using the proposed algorithm.

A diffuse sound field is composed of plane waves incident from all directions with equal probability and amplitude [69]. Using the scattering model previously introduced, we can determine the cross-correlation between the sound pressure at positions $\tilde{\mathbf{r}}$ and $\tilde{\mathbf{r}}'$ on the surface of a sphere, due to a unit amplitude plane wave with a random uniformly distributed direction of arrival (see Appendix B for derivation) [60]:

$$C(\tilde{\mathbf{r}}, \tilde{\mathbf{r}}', k) = \sum_{l=0}^{\infty} |b_l(k)|^2 (2l+1) \mathcal{P}_l(\cos \Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{r}}'}), \tag{3.23}$$

where $\Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{r}}'}$ is the angle between $\tilde{\mathbf{r}}$ and $\tilde{\mathbf{r}}'$. In the open sphere case, it is shown in Appendix B that this simplifies to the well-known spatial domain expression [69, 88, 118] $\mathrm{sinc}(k\,||\tilde{\mathbf{r}} - \tilde{\mathbf{r}}'||)$, where sinc denotes the unnormalized sinc function.

For the sound energy at a position $\tilde{\mathbf{r}}$ we substitute $\Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{r}}'} = 0$ and find $C(\tilde{\mathbf{r}}, \tilde{\mathbf{r}}, k) = \sum_{l=0}^{\infty} |b_l(k)|^2 (2l+1)$. According to SRA theory [69, 118], for frequencies above the Schroeder frequency [69] the energy of the reverberant sound field H_r is then given by [118]

$$\begin{aligned} \mathrm{E_s}\{|H_\mathrm{r}(\tilde{\mathbf{r}}, k)|^2\} &= \frac{1-\bar{\alpha}}{\pi A \bar{\alpha}} C(\tilde{\mathbf{r}}, \tilde{\mathbf{r}}, k) \\ &= \frac{1-\bar{\alpha}}{\pi A \bar{\alpha}} \sum_{l=0}^{\infty} |b_l(k)|^2 (2l+1), \end{aligned} \tag{3.24}$$

where $\mathrm{E_s}\{\cdot\}$ denotes spatial expectation, $\bar{\alpha}$ is the average wall absorption coefficient and A is the total wall surface area.

The above theoretical expression for the average reverberant energy can be compared to simulated results obtained using our method. We computed the spatial expectation using an average over 200 source-array positions, using the approach in Radlović et al. [88]: the array and source were kept in a fixed configuration (at a distance of 2 m from each other), which was then randomly rotated and translated. Both sources and

microphones were kept at least half a wavelength from the boundaries of the room, helping to ensure the diffuseness of the reverberant sound field [69]. The reverberant component H_r of the ATFs was computed by subtracting the direct path H_d from the simulated ATFs.

The room dimensions were equal to $6.4 \times 5 \times 4$ m, as in [88,110], such that the ratio of the dimensions was $(1.6 : 1.25 : 1)$, as recommended in [66, 88] to approximate a diffuse sound field. The reverberation time T_{60} was set to 500 ms, giving an average wall absorption coefficient of $\bar{\alpha} = 0.2656$. We simulated AIRs with a length of 4096 samples at a sampling frequency of 8 kHz. We considered frequencies from 300 Hz to 4 kHz, well above the Schroeder frequency of $2000\sqrt{\frac{0.5}{4 \cdot 5 \cdot 6.4}} = 125$ Hz, and the half-wavelength minimum distance is therefore 57 cm for a speed of sound of 343 m/s. We averaged the results over the 200 source-array positions and 32 quasi-equidistant receiver positions.

In Fig. 3.7, we plot the theoretical and simulated energy of H_r as a function of frequency, for two array radii (4.2 cm and 15 cm). We note that, except at low frequencies, there is a good match between the theoretical diffuse field energy expression we derived and the results obtained using our method. At lower frequencies, the theoretical equation overestimates the energy; we hypothesize that this is due to the reverberant sound field not being fully diffuse.

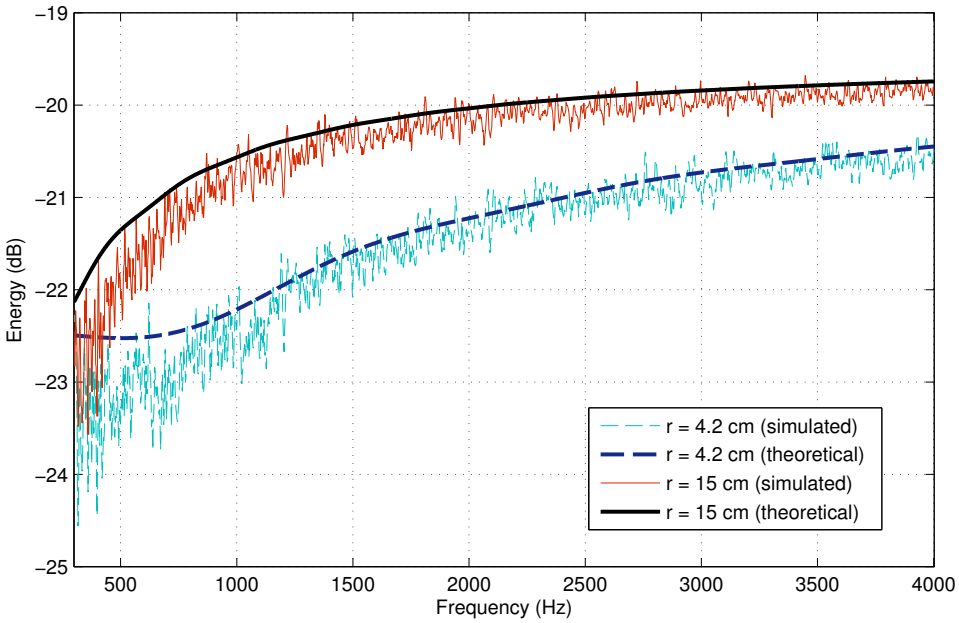


Figure 3.7: Theoretical and simulated reverberant sound field energy on the surface of a rigid sphere, as a function of frequency for two array radii. The simulated results are averaged over 200 source-array positions, all at least half a wavelength from the room boundaries.

3.4.2 Binaural interaural time and level differences

The topic of binaural sound and in particular head-related transfer functions (HRTFs) or head-related impulse responses (HRIRs) is of interest to researchers and engineers working on surround sound reproduction, who for example aim to reproduce spatial audio through a pair of stereo headphones. In addition, the psychoacoustic community is interested in the ability of the human brain to localize sound sources using only two ears.

Two binaural cues that contribute to sound source localization in humans are the interaural time difference (ITD) and the interaural level difference (ILD) [98]. The ITD measures the difference in arrival time of a sound at the two ears, and the ILD measures the level difference between the two ears. In this example, we study the long-term cues assuming the source signal is spectrally white. Therefore, we can compute the cues directly using the simulated ATFs.

We used the proposed method to simulate a simple HRTF by considering micro-

phones placed at locations on a rigid sphere corresponding to ear positions on the human head. Although real HRTFs vary from individual to individual, depending on the head, torso and pinnae, the main characteristics of the HRTF are also exhibited by a simple rigid sphere ATF [33]. The representation of HRTFs using spherical harmonics was studied in [8, 38].

Whereas HRTFs do not include the effects of reverberation, and as a result typically sound artificial and provide poor cues for the perception of sound source distance [102], the proposed method also allows for the inclusion of reverberation in HRIRs. In this case, they are then referred to as binaural room impulse responses (BRIRs). BRIRs are important for the analysis of the effects of reverberation on auditory perception, for example its impact on localization accuracy. Since rotational symmetry no longer necessarily holds once the room reflections are taken into account, the measurement of BRIRs must be done for every source-head position and orientation and is therefore very time-consuming. Simulating BRIRs allows us to more easily study the effects of early and late reflections on the binaural cues.

We begin by looking at ITDs in an anechoic environment, in order to illustrate the effect of the head in isolation. We compare simulated results to approximate theoretical results provided by a ray-tracing formula attributed to Woodworth & Schlosberg that looks at the distance travelled from the source to an observation point on the sphere, either in free-space if the observation point is on the near side of the sphere, or via a point of tangency if the observation point is on the far side [33].

The simulated results were obtained by generating HRIRs at a sampling frequency of 32 kHz, with a sphere radius of 8.75 cm and microphones placed at $(90°, 100°)$ (corresponding to the left ear) and $(90°, 260°)$ (corresponding to the right ear). The HRIRs were then band-pass filtered between 2.8 and 3.2 kHz[5]. The DOA was varied by ro-

[5]While the ray-tracing formula is frequency-independent, it has been shown [20] that ITDs actually exhibit some frequency dependence, and that because the ray-tracing concept applies to short wavelengths, this model yields only the high frequency time delay. Kuhn provides a more comprehensive discussion of this model and the frequency-dependence of ITDs [68]. It should be noted the simulation results in Fig. 3.8 are in broad agreement with Kuhn's measured results at 3.0 kHz.

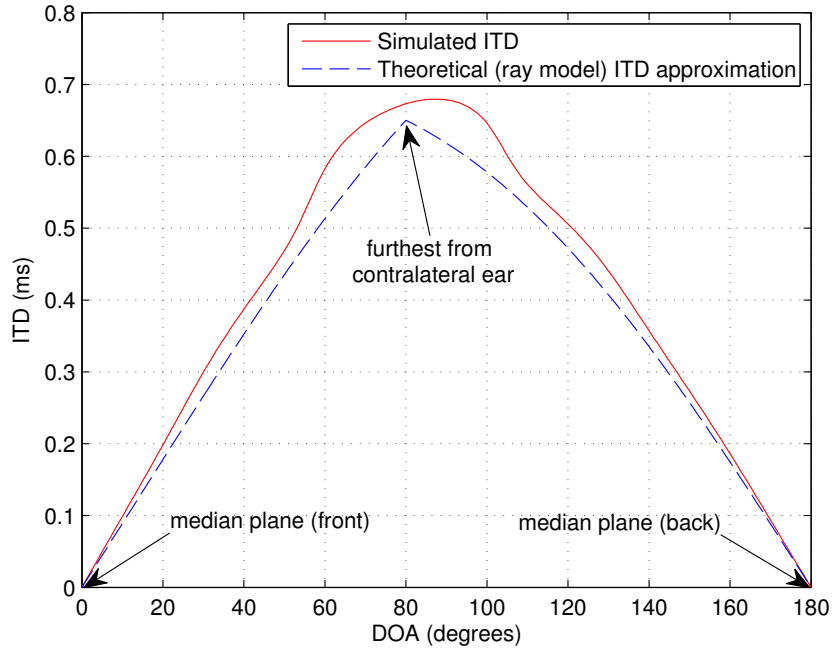


Figure 3.8: Comparison of ITDs as a function of source DOA, in simulation and using the theoretical ray model approximation. The simulated ITDs are based on HRIRs computed using our proposed algorithm in an anechoic environment.

tating the source around the sphere at a fixed distance of 1 m and inclination of 90°. The simulated ITD was computed by determining the time delay that maximized the interaural cross-correlation between the two simulated and band-pass filtered HRIRs. The cross-correlation was interpolated using a second-order polynomial in order to obtain sub-sample delays.

In Fig. 3.8 we plot the ITDs as a function of direction of arrival, where 0° corresponds to the median plane on the front side of the sphere and 180° corresponds to the median plane on the back side of the sphere. As expected, as the DOA increases from 0° to 80° and the source gets closer to the ipsilateral ear, the ITD increases monotonically until it reaches its maximum at 80°, at which point the source is furthest from the contralateral ear. The ITD then decreases from 80° to 180° as the source nears the median plane and gets closer to the contralateral ear. The response from 180° to 360° is not shown due to the symmetry about 180°. As we expect, our simulated results are reasonably close to the

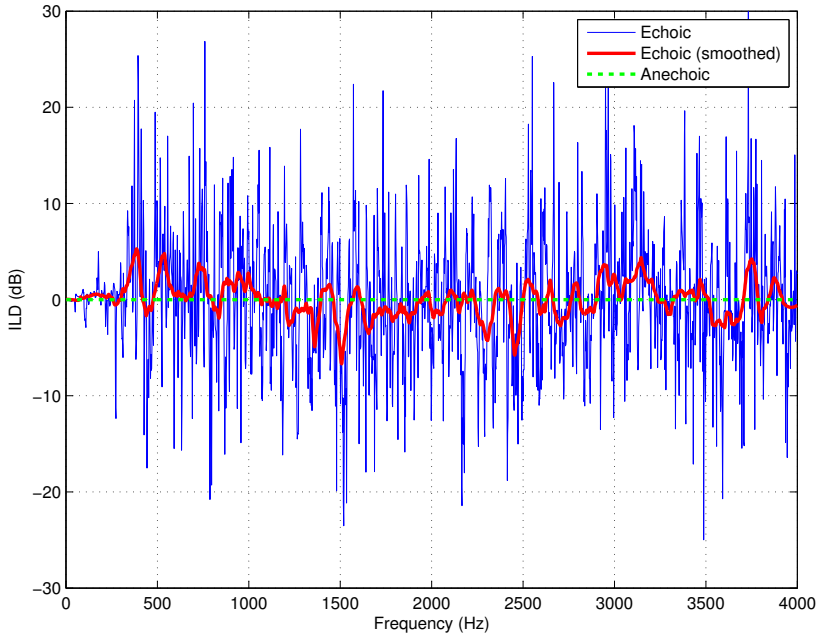


Figure 3.9: Comparison of ILDs in echoic and anechoic environments, with the sphere placed in the centre of the room and a DOA of 0°. The ILDs are based on HRTFs (anechoic) and BRIRs (echoic) computed using the proposed method.

theoretical ray-tracing results [33], with a difference of less than 70 µs.

Using the proposed method, we analyzed the ILDs in a reverberant environment under three scenarios: the sphere was either placed in the centre of the room with a DOA of 0° (where the source is equidistant from the two ears), or at a distance of approximately 0.5 m from one of the walls with DOAs of 0° and 100° (where the source is aligned with the left ear). In all three cases the source was placed at a distance of 1 m from the centre of the sphere. We chose a room size of 9 × 5 × 3 m with a reverberation time T_{60} of 500 ms, and simulated BRIRs with a length of 4096 samples at a sampling frequency of 8 kHz.

In Figs. 3.9, 3.10 & 3.11 we plot the ILDs for the three above cases, as well as the ILDs we would obtain in an anechoic environment, which are entirely due to scattering. The ILDs were computed by taking the difference in magnitude between the left ear response and the right ear response. A negative ILD therefore indicates that the magnitude of the ipsilateral ear response is lower than that of the contralateral ear response. The smoothed echoic ILDs were obtained using a Savitzky-Golay smoothing filter [99].

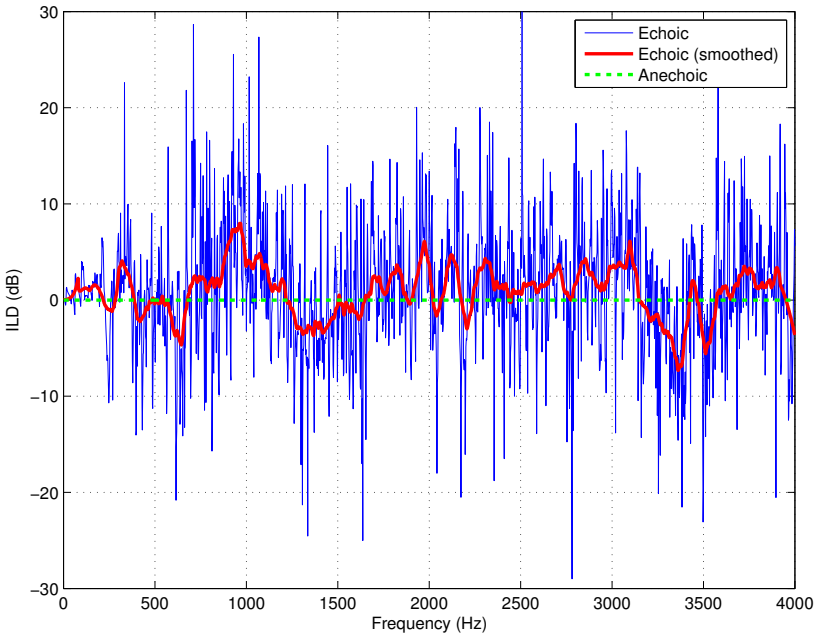


Figure 3.10: Comparison of ILDs in echoic and anechoic environments, with the sphere placed near a room wall and a DOA of 0°.

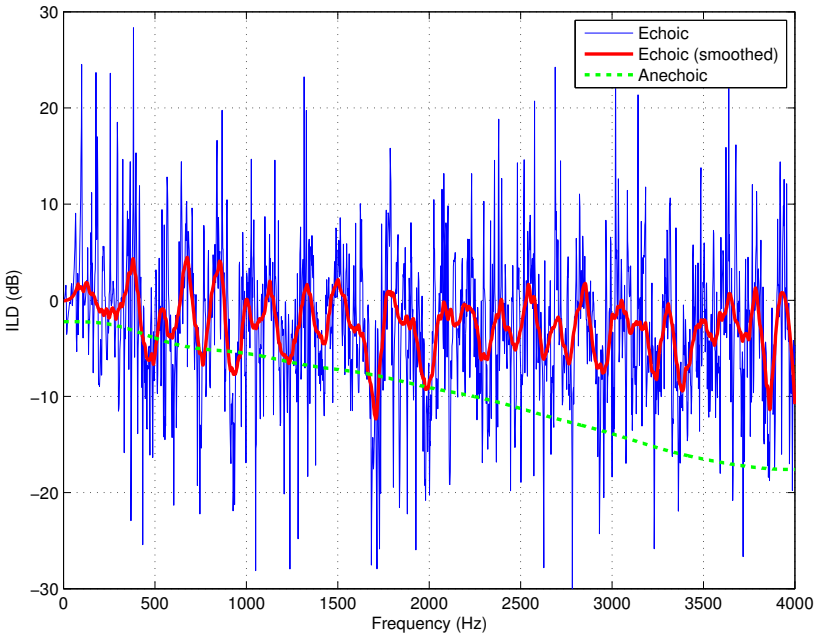


Figure 3.11: Comparison of ILDs in echoic and anechoic environments, with the sphere placed near a room wall and a DOA of 100°.

The main effect of reverberation we can observe is the introduction of random frequency-to-frequency variations; these are particularly obvious when most of the reverberant energy is diffuse, i.e. when the sphere is placed in the centre of the room (Fig. 3.9). Room reflections also increase the overall reverberant energy, particularly in the contralateral ear which receives less direct path energy, thus reducing the ILDs. This is especially noticeable when the contralateral ear is placed near a wall: the contralateral ear receives more energy than in the anechoic case and the ILD is therefore closer to zero (Fig. 3.11).

Placement of the sphere near a wall additionally introduces systematic distortions in the ILDs associated with the prominent early reflection from this wall. This is visible in Fig. 3.11 and most noticeably in Fig. 3.10.

All these effects have also been observed experimentally with a manikin by Shinn-Cunningham et al. [102]. The proposed algorithm is therefore an inexpensive way of predicting the effects of head movement and environmental changes (such as reverberation time) on HRTFs or BRIRs, without the need for more cumbersome experiments with head and torso simulators for example.

3.4.3 Mouth simulator

The principle of reciprocity can often be advantageously used in room acoustics measurements. The principle states that ATFs are symmetric in the coordinates of the sound source and the observation point: *"If we put the sound source at* $\mathbf{r}$*, we observe at point* $\mathbf{r}_0$ *the same sound pressure as we did before at* $\mathbf{r}$*, when the sound source was at* $\mathbf{r}_0$*"* [69]. We can apply this principle to ATF simulations, and use our method to generate the ATF between one or more sources on a sphere and a single omnidirectional microphone placed away from the sphere.

A specific application of this is a mouth simulator: we model the head as a rigid sphere (as in Section 3.4.2) of radius r_h, and the mouth as an omnidirectional point source placed on this rigid sphere. This is straightforwardly implemented in the proposed method

by replacing the source position with the microphone position $\tilde{\mathbf{r}}_{\text{mic}}$, the microphone position with the mouth position $\tilde{\mathbf{r}}_{\text{mouth}} = (r_{\text{h}}, \Omega_{\text{mouth}})$, and the array position with the head position, i.e.,

$$H(\tilde{\mathbf{r}}_{\text{mic}}|\tilde{\mathbf{r}}_{\text{mouth}}, k) = H(\tilde{\mathbf{r}} = \tilde{\mathbf{r}}_{\text{mouth}}|\tilde{\mathbf{r}}_{\text{s}} = \tilde{\mathbf{r}}_{\text{mic}}, k).$$

As a result we can simulate the ATF between a mouth on a head, and a single microphone in free space. Repeated use of the algorithm allows for multiple receivers.

Although more accurate modelling of the head and mouth is possible using finite element or boundary element methods for example, our algorithm is valuable for application to this problem due its comparative simplicity and the fact that, if desired, it can also take into account room reverberation. While the diameter of the mouth plays an important role in determining the filter characteristic of the vocal tract [30], we assume for the purposes of the scattering model that the mouth is a point source.

As an illustration of this application, Fig. 3.12 shows the energy of theATF between the mouth and a microphone as a function of microphone position at frequencies of 100 Hz and 3 kHz in an anechoic environment. The mouth was positioned on a sphere of radius 8.75 cm. Only two dimensions, x and y, are shown for brevity since the z dimension is identical to x and y. We observe that at 100 Hz there is no scattering and the radiation pattern is omnidirectional so that the sphere has little effect. At 3 kHz the effect of scattering starts to become more significant, and the energy at the back of the sphere is reduced while the energy at the front is increased. Finally the bright spot discussed in Section 3.2.3 is particularly apparent at the very back of the sphere in the bottom plot.

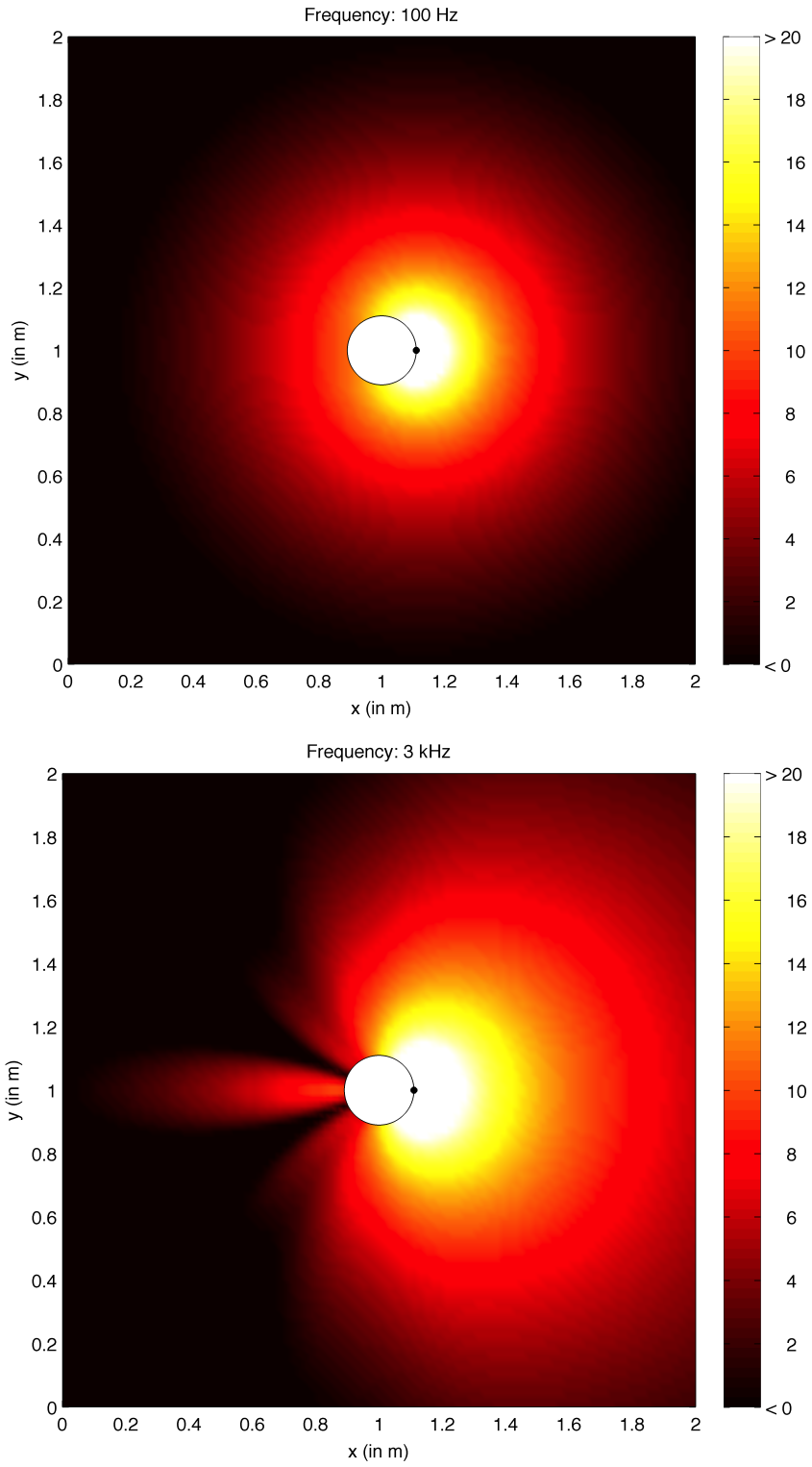


Figure 3.12: Sound energy radiation pattern (in dB) at 100 Hz (top) and 3 kHz (bottom). The mouth position is denoted by a black dot.

3.5 Conclusions

Spherical microphone arrays on a rigid baffle are of great interest currently. In order to analyze, work with and develop acoustic signal processing algorithms that make use of a spherical microphone array, a simulator is desired that can take into account the effects of the acoustic environment of the array as well as the scattering effects of the rigid spherical baffle. Accordingly, in this chapter a method was presented for the simulation of AIRs or ATFs for a rigid spherical microphone array in a reverberant environment.

We presented a scattering model used to model the rigid sphere, justifying its use with references to the literature, and provided an overview of the model's behaviour. We showed that the error with respect to the theoretical model can be controlled at the expense of increased computational complexity. Finally we provided a number of examples showing additional applications of this method.

Chapter 4

Spatial acoustic parameter estimation

Acoustic parameter estimation, the estimation of quantities which describe the sound field, is a major field of research within acoustic signal processing. Considerable research interest has focused on the estimation of parameters relating to sound sources, such as the number of active sound sources and their direction of arrival (DOA), initially inspired by work in wireless communications. It can also be useful to estimate room acoustic parameters, e.g., reverberation time, or parameters which relate to both the acoustic environment and the sound source, like the signal-to-diffuse energy ratio.

The estimation of certain acoustic parameters can provide additional *a priori* information to speech enhancement algorithms, thereby improving their performance. In this chapter, we propose methods for estimating two such parameters: theDOA of a source, for both the static (Section 4.1) and moving (Section 4.2) cases, and the signal-to-diffuse ratio or diffuseness (Section 4.3).

4.1 Direction of arrival estimation

In this section, we seek to perform two-dimensional DOA estimation (azimuth and inclination)[1] using spherical microphone arrays, which is useful in applications such as beamforming (see Chapter 5), noise source identification (in vehicles or aircraft, for example), or automatic camera steering.

One-dimensional DOA estimation (azimuth only) has been widely studied, using time difference of arrival (TDOA)-based methods (such as GCC-PHAT), subspace-based methods (ESPRIT, MUSIC), or steered response power (SRP). MUSIC and ESPRIT have also been generalized to two dimensions and extended to the spherical harmonic domain (SHD) [65,113], although they are typically not robust to reverberation, and both MUSIC and SRP are computationally inefficient due to the need for an exhaustive search. Additionally TDOA-based methods are unsuitable for practical spherical microphone arrays with a small radius, due to the insufficient spacing between microphones.

In this work, we propose a two-dimensional DOA estimation method for spherical microphone arrays, based on a pseudointensity vector that indicates the direction of the active sound source. This vector is calculated using only the zero- and first-order eigenbeams. We compare the proposed method to a SHD implementation of the SRP method which is commonly used in the spatial domain.

This work relates to previous intensity vector–based DOA estimation work in the field of Directional Audio Coding (DirAC) [5], although the pseudointensity vector is calculated using eigenbeams, while the intensity vector is computed using the Ambisonic B-format signals, which are often measured directly (using an omnidirectional microphone and three dipole microphones) or with a three or four omnidirectional microphone grid. The eigenbeams we use for DOA estimation are computed using

Portions of this work were first published in the *Proceedings of the 18th European Signal Processing Conference (EUSIPCO-2010)* [57] in 2010, published by EURASIP.

[1] We refer to this problem as *two-dimensional* DOA estimation rather than *three-dimensional source localization* to reflect the fact that we do not estimate the source-array distance. The source position is not, however, confined to a two-dimensional space.

all of the microphones in a spherical array, of which there are typically a few dozen, thus providing more robustness to noise that is incoherent in the SHD (either spatially incoherent noise, or diffuse noise, as shown in Section 4.3.2). An earlier version of this work was previously published in [57].

4.1.1 Spherical harmonics

Consider a sound pressure field at a point $\mathbf{r} = (r, \Omega) = (r, \theta, \phi)$ (in spherical polar coordinates, with inclination θ and azimuth ϕ), denoted by $P(k, \mathbf{r})$, where k is the wavenumber.

The spherical Fourier transform of $P(k, \mathbf{r})$ is denoted by $P_{lm}(k)$, as defined in (2.4). In a Q microphone system with spherical coordinates $\mathbf{r}_q = (r_q, \Omega_q), q = 1, \ldots, Q$, we must approximate the integral in (2.4) with a sum:

$$P_{lm}(k) \approx \sum_{q=1}^{Q} \mathfrak{g}_{q,lm} P(k, \mathbf{r}_q). \tag{4.1}$$

The number of microphones Q and the quadrature weights $\mathfrak{g}_{q,lm}$ must be chosen such that (4.1) is a sufficiently accurate approximation of (2.4), as explained in Section 2.3.

4.1.2 Direction of arrival estimation using the steered response power

As a baseline for comparison, we now present a SHD equivalent of a conventional spatial domain DOA estimation method; namely, computing a map of the SRP, which is obtained by steering a beamformer in various directions and determining the output power. The DOA is then obtained by finding the direction with the highest power. In order to produce this acoustic map using eigenbeams, we first introduce the theory of beamforming in the SHD.

4.1.2.1 Beamforming

The eigenbeams $P_{lm}(k)$ that result from the spherical Fourier transform can be interpreted as individual sensors in the classical sensor array processing framework. It is important to note that the directivity pattern of the eigenbeams is frequency-invariant, while each magnitude response depends on the order l.

Once we have computed the eigenbeams, we can synthesize an arbitrary beam pattern by applying a modal beamformer. In the same way that the output of a spatial domain beamformer can be expressed as a weighted sum of the spatial domain input signals, the output of a modal beamformer can be expressed as a weighted sum of the eigenbeams[2], i.e.,

$$Z(k) = \sum_{l=0}^{L} \sum_{m=-l}^{l} W_{lm}^{*}(k) P_{lm}(k), \tag{4.2}$$

where L is the array order and $W_{lm}(k)$ are the SHD beamforming weights.

It is often sufficient to use a beam pattern which is rotationally symmetric around the look direction Ω_{u}, in which case the weights can be expressed as [77]

$$W_{lm}^{*}(k, \Omega_{\mathrm{u}}) = \frac{d_l(k)}{b_l(k)} Y_{lm}(\Omega_{\mathrm{u}}), \tag{4.3}$$

where $d_l(k)$ allows us to control the beam pattern, Y_{lm} is the spherical harmonic of order l and degree m as defined in (2.5), and $b_l(k)$ is the mode strength, as defined in Section 2.4. While the above interpretation has some practical advantages, it should be noted that the inverse spherical Fourier transform given by (2.3) is done implicitly as it is incorporated into the beamformer weights.

[2] In practice, the acquired pressure signals in the time domain are normally transformed to the short-time Fourier transform domain, such that they depend on the time index, although this dependency is omitted for brevity.

By combining (4.2) and (4.3) and reorganizing the terms we obtain

$$Z(k,\Omega_\mathrm{u}) = \sum_{l=0}^{L}\sum_{m=-l}^{l}\frac{d_l(k)}{b_l(k)}Y_{lm}(\Omega_\mathrm{u})P_{lm}(k) \tag{4.4a}$$

$$= \sum_{l=0}^{L}\frac{d_l(k)}{b_l(k)}\sum_{m=-l}^{l}Y_{lm}(\Omega_\mathrm{u})P_{lm}(k). \tag{4.4b}$$

From (4.4b) it can be seen that the output of the beamformer can be computed in two steps. In the first step (the inner summation) the beamformer is steered to the look direction Ω_u. In the second step (the outer summation) the beam pattern is synthesized.

We take advantage of the orthonormality of the spherical harmonics in (2.6) and choose weights $\mathfrak{g}_{q,lm}$ given by

$$\mathfrak{g}_{q,lm} = \frac{4\pi}{Q}Y_{lm}^*(\Omega_q), \tag{4.5}$$

which makes the approximation in (4.1) exact if $Q \geq (L+1)^2$ and the microphones are equally spaced on the sphere. For non-trivial microphone configurations, it is not possible for the microphones to be perfectly equidistant, therefore a small error is involved. By substituting the expression for the weights $\mathfrak{g}_{q,lm}$ in (4.5) into (4.1) we obtain

$$P_{lm}(k) \approx \frac{4\pi}{Q}\sum_{q=1}^{Q}Y_{lm}^*(\Omega_q)P(k,\mathbf{r}_q), \tag{4.6}$$

and substituting this expression into the beamformer output $Z(k,\Omega_\mathrm{u})$ expression in (4.4b), choosing $d_l(k) = 1$ (which yields the plane-wave decomposition beamformer from Sec. 2.5 with maximum directivity [95]), we find an expression relating the beamformer output $Z(k,\Omega_\mathrm{u})$ to the measured pressure signals $P(k,\mathbf{r}_q)$:

$$Z(k,\Omega_\mathrm{u}) \approx \frac{4\pi}{Q}\sum_{l=0}^{L}\frac{1}{b_l(k)}\sum_{m=-l}^{l}Y_{lm}(\Omega_\mathrm{u})\sum_{q=1}^{Q}Y_{lm}^*(\Omega_q)P(k,\mathbf{r}_q). \tag{4.7}$$

The theoretical beamformer output can be predicted by assuming a single active sound source and far-field conditions, in which case the wavefront impinging on a

spherical array of radius r can be assumed to be planar, and if we denote its arrival direction as Ω_0, we can write $P_{lm}(k)$ as [89]

$$P_{lm}(k) = A(k) b_l(k) Y_{lm}^*(\Omega_0), \tag{4.8}$$

where $A(k)$ is the wave amplitude. Substituting (4.8) in (4.4b), and choosing $d_l(k) = 1$, we obtain

$$\begin{aligned} Z(k, \Omega_\mathrm{u}) &= \sum_{l=0}^{L} \frac{1}{b_l(k)} \sum_{m=-l}^{l} Y_{lm}(\Omega_\mathrm{u}) A(k) b_l(k) Y_{lm}^*(\Omega_0) && \text{(4.9a)} \\ &= A(k) \sum_{l=0}^{L} \sum_{m=-l}^{l} Y_{lm}(\Omega_\mathrm{u}) Y_{lm}^*(\Omega_0) && \text{(4.9b)} \\ &= \begin{cases} \dfrac{A(k)(L+1)^2}{4\pi} & \text{if } \Omega_0 = \Omega_\mathrm{u}, \\ \dfrac{A(k)(L+1)}{4\pi(\cos\Theta - 1)} \left[\mathcal{P}_{L+1}(\cos\Theta) - \mathcal{P}_L(\cos\Theta)\right] & \text{otherwise,} \end{cases} && \text{(4.9c)} \end{aligned}$$

where Θ is the angle between Ω_0 and Ω_u. The last step in the derivation is explained in [89]. The beamformer output $Z(k, \Omega_\mathrm{u})$ reaches its maximum when $\Theta = 0$ [89], i.e., when the look direction Ω_u is equal to the arrival direction Ω_0, as desired. We normalize the beamformer output with respect to its value for $\Theta = 0$, and plot it as a function of Θ in Fig. 4.1. We see that as L increases, the distribution of $Z(k, \Omega_\mathrm{u})$ narrows around $\Theta = 0$, tending towards a delta function for $L \to \infty$ [89].

4.1.2.2 Steered response power map

An acoustic map can be computed and depicted in different ways. Here we choose to compute the power corresponding to the output of a beamformer steered in different directions. The direction with the highest power provides an estimate of the location of the sound source. The resolution of the acoustic map depends on the directivity pattern of the beamformer (which in turn depends on the array order L), and on the number of beams for which the power is measured.

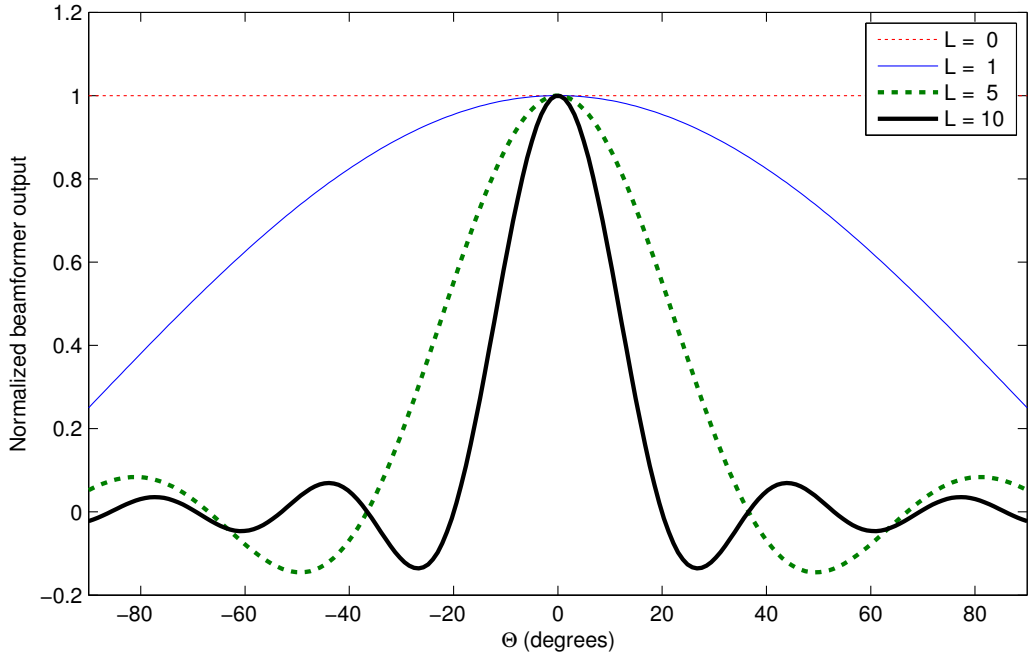


Figure 4.1: Normalized beamformer output as a function of the beamformer order L and Θ, the angle between the beamformer look direction and the DOA.

From the expression for the beamformer output as a function of the look direction Ω_u, we can compute a power map $\mathcal{M}(\Omega_u)$ for each direction Ω_u by averaging $|Z(k, \Omega_u)|^2$ across a number of discrete wavenumber values (denoted by $\dot{k}$), forming a set $\mathcal{K}$, i.e.,

$$\mathcal{M}(\Omega_u) = \sum_{\dot{k} \in \mathcal{K}} \beta_Z(\dot{k}) \left| Z(\dot{k}, \Omega_u) \right|^2, \tag{4.10}$$

where $\beta_Z(\dot{k})$ is a weighting function which allows us to, for example, ignore all beams below a certain frequency, which are likely to contain low frequency noise and little speech, or to apply an A-weighting function. We can also smooth the map over multiple time frames, depending on the time resolution which is desired for the DOA estimates.

Assuming a single active source, the source DOA Ω_s is then given by the direction with maximum power:

$$\Omega_s = \arg\max_{\Omega_u} \mathcal{M}(\Omega_u). \tag{4.11}$$

4.1.3 Direction of arrival estimation using the pseudointensity vector

4.1.3.1 Motivation

While intuitively simple, the SRP method is computationally complex: as the function $\mathcal{M}(\Omega_u)$ is non-convex, we must steer a beam in many directions to determine which direction has the highest power, and hence where the sound source is likely to be located. We now present a novel alternative method for DOA estimation with low computational complexity, based on pseudointensity vectors.

In acoustics, sound intensity is a measure of the flow of sound energy through a surface per unit area, in a direction perpendicular to this surface. The idea of a pseudointensity vector is inspired by the concept of intensity vectors, defined as [27]

$$\mathcal{I} = \frac{1}{2}\Re\left\{P^* \cdot \mathbf{v}\right\}, \tag{4.12}$$

where P is the sound pressure, $\mathbf{v} = [V_x, V_y, V_z]^T$ is the particle velocity vector, and $\Re\{\cdot\}$ denotes the real part of a complex number. For a single plane wave, the particle velocity vector is given by [26, p. 31]

$$\mathbf{v} = -\frac{P}{\rho_0 c}\mathbf{u}, \tag{4.13}$$

where c is the speed of sound in the medium, ρ_0 is the ambient density, and $\mathbf{u}$ is a unit vector pointing towards the acoustic source. Consequently, the intensity vector points in the direction opposite to the vector $\mathbf{u}$.

The intensity vector corresponds to the magnitude and direction of the transport of acoustical energy, indicating its utility for determining the DOA of a sound wave. Unfortunately in practice it is difficult to measure particle velocity, although attempts have been made using vibrating surfaces and accelerometers, or more successfully, using the finite difference method with dual-microphone arrays [27]. More recently particle velocity has been measured with a micromachined transducer, the Microflown [28]. In order to be able to use only one type of sensor, we would like to compute the intensity

vector using a spherical microphone array.

4.1.3.2 Definition

We propose a pseudointensity vector $\mathbf{I}(k)$ which is conceptually similar to an intensity vector, but is calculated using the zero- and first-order eigenbeams $P_{lm}(k)$ ($l = 0, 1$), and is defined as follows:

$$\mathbf{I}(k) = \frac{1}{2}\Re\left\{\left(\frac{P_{00}(k)}{b_0(k)}\right)^* \begin{bmatrix} P_x(k) \\ P_y(k) \\ P_z(k) \end{bmatrix}\right\}, \tag{4.14}$$

where the first term, $(P_{00}(k)/b_0(k))^*$ is the complex conjugate of the omnidirectional sound pressure signal, and the second term corresponds to the particle velocity vector in (4.12). The components $P_x(k)$, $P_y(k)$ and $P_z(k)$ of this vector are dipoles steered in the opposite direction to the x, y and z axes [see Fig. 4.2 for a plot of the beam pattern of $P_x(k)$]. These dipoles approximate the pressure gradient, which is proportional to the particle velocity [75, 119]. Since we are only interested in the pseudointensity vector's direction, the scale factor $(\rho_0 c)^{-1}$ is omitted here.

In order to form the beams $P_x(k)$, $P_y(k)$ and $P_z(k)$, we make use of the available eigenbeams $P_{1(-1)}(k)$, $P_{10}(k)$ and $P_{11}(k)$. This can be done by forming a linear combination of rotated eigenbeams, i.e., implementing a plane-wave decomposition beamformer as defined in (2.14):

$$P_a(k) = \frac{1}{b_1(k)} \sum_{m=-1}^{1} \alpha_{a,m}\, P_{1m}(k), \quad a \in \{x, y, z\}, \tag{4.15}$$

where the $b_1(k)$ factor is required to make the beam patterns independent of the wavenumber.

To rotate each of the eigenbeams in the appropriate direction (θ_r, ϕ_r), we multiply

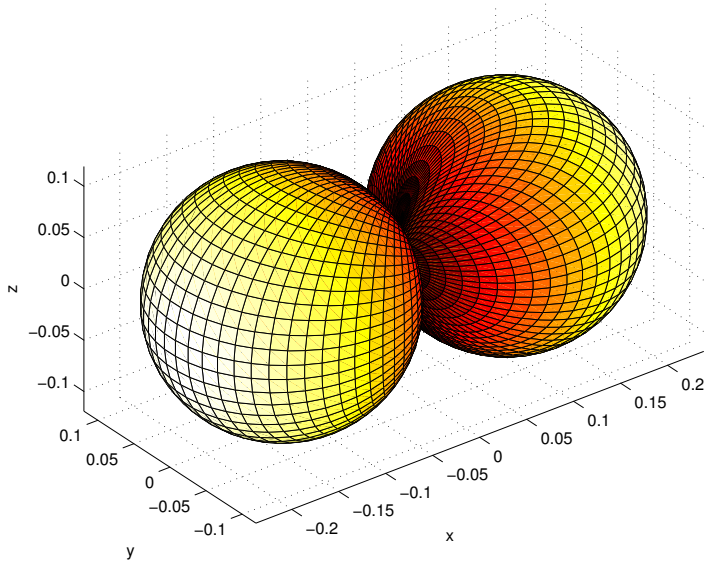


Figure 4.2: Beam pattern of the beam P_x, aligned to the x-axis: $|\alpha_{x,(-1)} Y_{1(-1)}(\theta,\phi) + \alpha_{x,0} Y_{10}(\theta,\phi) + \alpha_{x,1} Y_{11}(\theta,\phi)|$.

them by the spherical harmonics $Y_{1m}(\theta_\mathrm{r},\phi_\mathrm{r})$. We therefore require:

$$\alpha_{x,m} = Y_{1m}(\pi/2,\pi), \tag{4.16a}$$

$$\alpha_{y,m} = Y_{1m}(\pi/2,-\pi/2), \tag{4.16b}$$

$$\alpha_{z,m} = Y_{1m}(\pi,0). \tag{4.16c}$$

The beam pattern of P_x, which is aligned to the x-axis, is shown as an example in Fig. 4.2.

4.1.3.3 Direction of arrival estimation

The pseudointensity vector is calculated for every discrete wavenumber; for every time instant we therefore have a number of vectors which point in slightly different directions. While they provide an approximate location for the sound source, some averaging is necessary to locate it more precisely. The intensity vector averaged across the discrete

wavenumbers $\dot{k}$ forming a set $\mathcal{K}$ is given by

$$\mathbf{I} = \sum_{\dot{k} \in \mathcal{K}} \beta_I(\dot{k}) \mathbf{I}(\dot{k}), \tag{4.17}$$

where $\beta_I(\dot{k})$ is a weighting function similar to $\beta_Z(\dot{k})$ in (4.10). Note that even with $\beta_I(\dot{k}) = 1$, $\forall \dot{k}$ we are implicitly giving a higher weight to the intensity vectors with the highest norm, which we consider to be more reliable for DOA estimation. If DOA estimates are required for every time and frequency instant, the wavenumber dependent pseudointensity vector $\mathbf{I}(k)$ can be used, at the expense of reduced accuracy.

An estimate of the unit vector $\mathbf{u}$ pointing in the direction of the sound source, as in (4.13), is given by normalizing the pseudointensity vector, i.e.,

$$\hat{\mathbf{u}} = -\frac{\mathbf{I}}{\|\mathbf{I}\|_2}, \tag{4.18}$$

where $\|\cdot\|_2$ denotes the ℓ-2 vector norm. When multiple time observations are available, one can additionally smooth $\mathbf{I}$ or $\hat{\mathbf{u}}$ over time.

4.1.4 Computational complexity

The pseudointensity method requires the computation of the four zero- and first-order eigenbeams, and three weighted averages $P_x(k)$, $P_y(k)$ and $P_z(k)$ of these eigenbeams. The SRP method, on the other hand, requires us to compute these eigenbeams (and potentially more eigenbeams if $L > 1$), and additionally steer beams in many directions as shown in (4.4).

A fair comparison of these two methods would therefore be to compute the SRP with only three beams, however for this number of beams it is impossible to obtain a reasonable DOA estimate from the SRP. As we will see in Section 4.1.5, to obtain accuracy of the same order as the pseudointensity vector method, we must steer thousands of beams.

In practice, however, it is not efficient to steer this many beams indiscriminately in all directions: a coarse grid approach can be taken at first, to determine the DOA within ±30°, for example, and we can then apply a finer grid to the area of interest, thus reducing the amount of unnecessary detail in directions where the acoustic source cannot be located (based on the results of the first search).

4.1.5 Performance evaluation

In order to evaluate and compare the performance of these two DOA estimation methods, we calculate the angle ϵ between a unit vector pointing in the correct direction $\mathbf{u}$, and a unit vector $\hat{\mathbf{u}}$ pointing in the direction estimated by either of the two methods, as in [15]. The angle ϵ is then given by

$$\epsilon = \cos^{-1}(\mathbf{u}^{\mathrm{T}}\hat{\mathbf{u}}). \tag{4.19}$$

4.1.5.1 Using simulated data

In order to objectively evaluate the accuracy of the pseudointensity vector DOA estimation method, we must generate pseudointensity vectors in a simulated environment where the true source positions are known precisely. We achieve this by simulating impulse responses with SMIRgen [54], an acoustic impulse response (AIR) simulator for spherical microphone arrays based on the algorithm presented in Chapter 3.

For these simulations we placed a $Q = 32$ microphone array with radius 4.2 cm near the centre of an acoustic space with dimensions $10 \times 8 \times 12$ m in which a single source was present. The source signal consisted of a white Gaussian noise sequence. We processed the signals in the short-time Fourier transform (STFT) domain with a sampling frequency of 8 kHz and a frame length of 64 ms with a 50% overlap. We averaged the acoustic map and pseudointensity vectors over 5 time frames, i.e., 192 ms of data. We used the same number of eigenbeams for the SRP as for the pseudointensity vector, i.e., we chose the limit $L = 1$. We did not apply any weighting in (4.10) and (4.17), that is, we set $\beta_Z(\dot{k}) = \beta_I(\dot{k}) = 1, \forall \dot{k}$. We added spatio-temporally white Gaussian noise

to the individual microphone signals in order to obtain an input signal-to-incoherent-noise ratio (iSINR) of 20 dB at the microphone closest to the source, i.e., the microphone with the highest iSINR.

In the first simulation, the reverberation time T_{60} was varied between 0 (anechoic room) and 600 ms while the source-array distance was fixed at 2.5 m. The room boundary reflection coefficients were computed from the desired reverberation times using Sabin-Franklin's formula [86]. With such a configuration, reverberation times between 300 and 600 ms corresponded to direct-to-reverberant energy ratios between approximately 10 and 0 dB. In the second simulation the source-array distance ranged between 1 and 3 m while the reverberation time was fixed at 450 ms.

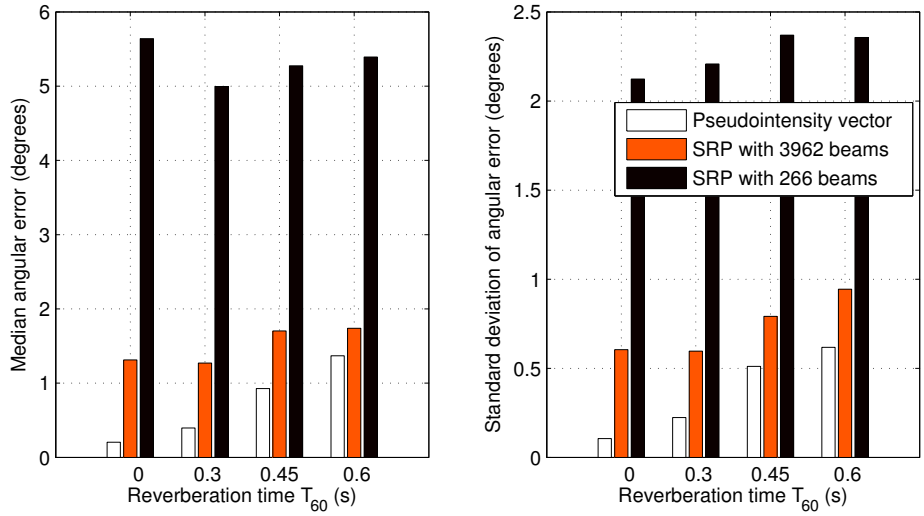
A statistical analysis of the results of these simulations is shown in Fig. 4.3, based on Monte Carlo simulations with 100 runs. For each run a new DOA was randomly selected from a uniform angular distribution around the sphere. The accuracy of the pseudointensity vector method is significantly higher than that of the SRP method with a small number of beams (266). For a larger number of beams (3962), the pseudointensity vector method still outperforms the SRP method, but by a smaller margin. This is still the case even as the source-array distance increases above 2 m and the reverberation time increases above 450 ms.

As expected, the accuracy of the proposed method increases as the source-array distance and reverberation time decrease, since both these changes lead to an increase in the direct-to-reverberant energy ratio. The robustness of the proposed method to reverberation is due to the fact that the reverberation is mostly diffuse, and therefore causes little bias in the DOA estimates once they have been averaged over frequency (and optionally over time).

4.1.5.2 Using spherical microphone array measurements

To experimentally test our proposed method, we measured a sound field using an em32 *Eigenmike* from mh acoustics, which is a commercially available spherical microphone

(a) Angular errors as a function of reverberation time T_{60} for a source-array distance of 2.5 m.



(b) Angular errors as a function of source-array distance for a reverberation time of 450 ms.

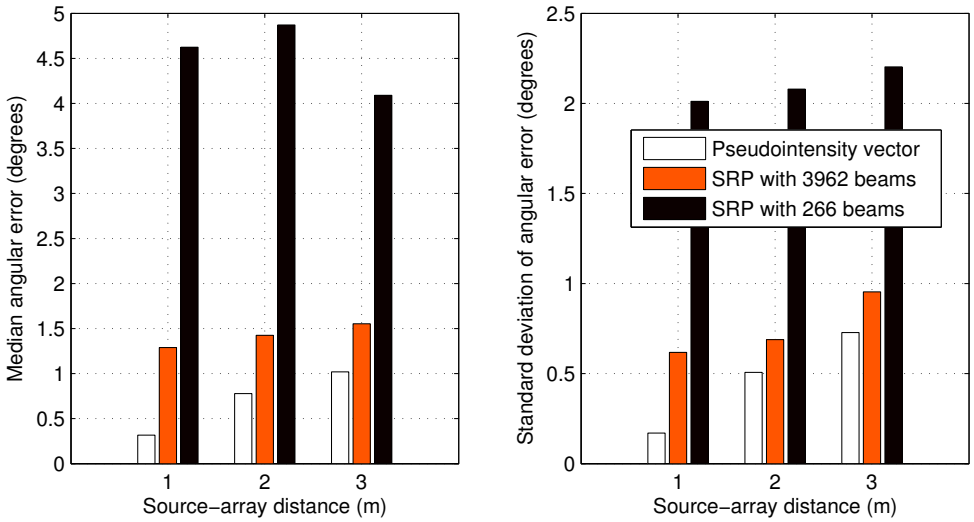


Figure 4.3: Median and standard deviation of the angular errors for the SRP and pseudointensity vector methods, as a function of reverberation time (a) and source-array distance (b). In (a) the source-array distance is 2.5 m and in (b) the reverberation time is 450 ms; both of these conditions ensure that the direct-to-reverberant energy ratio remains above 0 dB.

array of radius $r = 4.2$ cm with $Q = 32$ microphones. Measurements were taken in a room with dimensions $2.9 \times 2.7 \times 3.3$ m with a reverberation time of approximately 300 ms. The source signal consisted of 2 s of white Gaussian noise. We again chose $L = 1$.

Unfortunately as it was not possible to take precise measurements of the true DOAs, a quantitative assessment of the accuracy of the two methods would not be meaningful, however for illustrative purposes Fig. 4.4 shows a power map obtained using the SRP method, and Fig. 4.5 is a plot of the azimuths and inclinations of the DOAs obtained using the proposed method, for a source located at approximately $(87°, 36°)$. In Fig. 4.5 we note a cluster of DOA estimates centred around the correct DOA, and Fig. 4.4 confirms that the direction of highest power corresponds to this same DOA.

4.1.6 Conclusion

The pseudointensity vector offers the possibility of fast DOA estimation without the computational complexity of steering beams in all directions. Furthermore, the results it yields are highly accurate when compared to the SRP method with a viable number of beams (266): in typical environments, the median error is around 1°, as opposed to 4–5° with the SRP method.

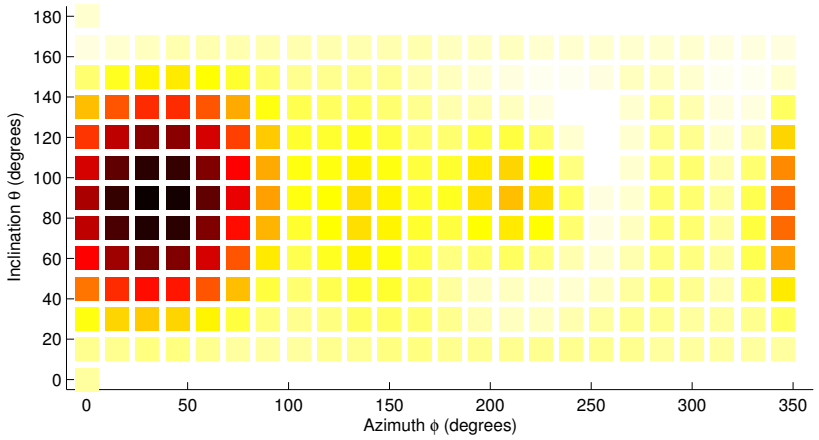


Figure 4.4: Power map for a source at approximately $(87°, 36°)$, with 266 beams. The darkest areas correspond to the beams with highest power.

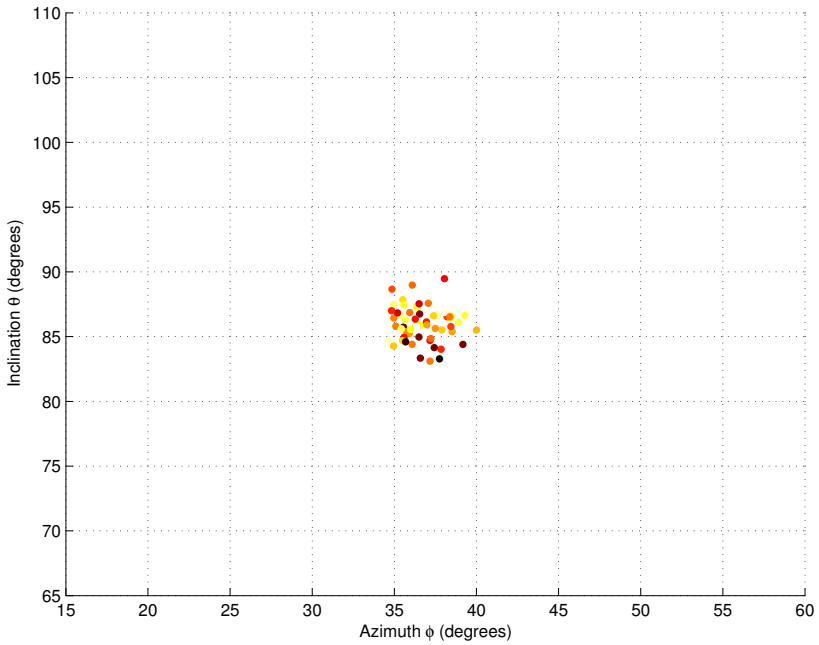


Figure 4.5: Plot of DOAs estimated using the proposed method at each time frame, for a source at approximately $(87°, 36°)$. The darkness of the data points indicates the norm of the corresponding vector.

4.2 Source tracking

In Section 4.1, we proposed a low computational complexity DOA estimation method based on eigenbeams. The eigenbeams were used to compute a pseudointensity vector indicating the DOA of the source. In this section, we use a similar eigenbeam-based method to estimate the particle velocity vector, which can also be used for DOA estimation [82]. An earlier version of this work was previously published in [58].

The source tracking is performed using an adaptive principal component analysis (PCA) of the acoustic particle velocity vector, and is robust to both noise and reverberation. The low complexity of the proposed method is crucial for real-time tracking applications. For plane wave incidence, the particle velocity vector points from the acoustic source to the receiver; we therefore assume far-field conditions, a point source and point sensors.

4.2.1 Problem formulation

4.2.1.1 Particle velocity vector

Let $\mathbf{u}$ be a unit vector pointing from the array towards an acoustic source. Assuming plane wave incidence, the acoustic particle velocity vector $\mathbf{s}$ at a discrete time instant n is given by [26, p. 31]

$$\mathbf{s}(n) = -\frac{p(n)}{Z_0}\mathbf{u}(n), \tag{4.20}$$

where p is the sound pressure and $Z_0 = \rho_0 c$ is the characteristic acoustic impedance of air.

Unfortunately, while the particle velocity vector can be computed using pressure measurements from a spherical microphone array, the resulting vector is corrupted by both noise and reverberation. We seek to mitigate these effects and accurately track the source DOA, i.e., the vector $\mathbf{u}$, by applying a beamformer to the noisy particle velocity

Portions of this work were first published in the *Proceedings of the IEEE Asilomar Conference on Signals, Systems, and Computers* [58] in 2010. © 2010 IEEE.

vector estimates.

4.2.1.2 Maximum signal-to-noise ratio beamforming

Let $\mathbf{v}(n) = [v_x(n), v_y(n), v_z(n)]^{\mathrm{T}}$ be the noisy input signal, a time-varying particle velocity vector. The noise is modelled by a term $\mathbf{e}(n) = [e_x(n), e_y(n), e_z(n)]^{\mathrm{T}}$ that can include both ambient noise and room reverberation. The desired signal $\mathbf{s}$ and noise signal $\mathbf{e}$ are assumed to be mutually uncorrelated; the reflections due to reverberation are therefore assumed to be diffuse. Our signal model is then given by

$$\begin{aligned}\mathbf{v}(n) &= \mathbf{s}(n) + \mathbf{e}(n) \\ &= -\frac{p(n)}{Z_0}\,\mathbf{u}(n) + \mathbf{e}(n).\end{aligned} \tag{4.21}$$

We can apply a time-varying spatial weighting vector $\mathbf{w}(n)$ to the input signal $\mathbf{v}(n)$, and sum the resulting three signals, to obtain an output signal $z(n)$ (the beamformer output):

$$\begin{aligned}z(n) &= \mathbf{w}^{\mathrm{T}}(n)\,\mathbf{v}(n) \\ &= \mathbf{w}^{\mathrm{T}}(n)\,\mathbf{s}(n) + \mathbf{w}^{\mathrm{T}}(n)\,\mathbf{e}(n) \\ &= -\frac{p(n)}{Z_0}\mathbf{w}^{\mathrm{T}}(n)\,\mathbf{u}(n) + \mathbf{w}^{\mathrm{T}}(n)\mathbf{e}(n).\end{aligned} \tag{4.22}$$

The signal-to-noise ratio (SNR) at the output of the beamformer can be defined as[3]

$$\begin{aligned}\mathrm{oSNR}(\mathbf{w}) &= \frac{\mathrm{E}\left\{\left[\mathbf{w}^{\mathrm{T}}\,\mathbf{s}\right]\left[\mathbf{w}^{\mathrm{T}}\,\mathbf{s}\right]^{\mathrm{T}}\right\}}{\mathrm{E}\left\{\left[\mathbf{w}^{\mathrm{T}}\,\mathbf{e}\right]\left[\mathbf{w}^{\mathrm{T}}\,\mathbf{e}\right]^{\mathrm{T}}\right\}} \\ &= \frac{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}_{\mathbf{s}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}_{\mathbf{e}}\mathbf{w}},\end{aligned} \tag{4.23}$$

where $\boldsymbol{\Phi}_{\mathbf{s}} = \mathrm{E}\{\mathbf{s}\mathbf{s}^{\mathrm{T}}\}$ is the covariance matrix of the desired signal and $\boldsymbol{\Phi}_{\mathbf{e}} = \mathrm{E}\{\mathbf{e}\mathbf{e}^{\mathrm{T}}\}$ is

[3] The dependency on n is omitted for brevity.

the covariance matrix of the noise. As the desired signal $\mathbf{s}$ and the noise signal $\mathbf{e}$ are mutually uncorrelated, the covariance matrix of the input signal $\mathbf{v}$ can be expressed as $\mathbf{\Phi}_\mathbf{v} = \mathbf{\Phi}_\mathbf{s} + \mathbf{\Phi}_\mathbf{e}$ and we can express the variance of the output z as

$$\begin{aligned}\sigma_z^2 &= \mathbf{w}^\mathrm{T}\mathbf{\Phi}_\mathbf{v}\mathbf{w} \\ &= \mathbf{w}^\mathrm{T}\left[\mathbf{\Phi}_\mathbf{s} + \mathbf{\Phi}_\mathbf{e}\right]\mathbf{w} \\ &= \mathbf{w}^\mathrm{T}\mathbf{\Phi}_\mathbf{s}\mathbf{w} + \mathbf{w}^\mathrm{T}\mathbf{\Phi}_\mathbf{e}\mathbf{w}. \end{aligned} \tag{4.24}$$

The beamformer with weights $\mathbf{w}$ that maximizes the output SNR $\mathrm{oSNR}(\mathbf{w})$ is known as a maximum SNR beamformer. This is equivalent to determining the principal component of the data set comprising the noisy observations of the particle velocity vector.

Let us now assume spherically white noise such that

$$\mathbf{\Phi}_\mathbf{e} = \sigma_e^2\mathbf{I}_{3\times 3}, \tag{4.25}$$

where $\mathbf{I}_{3\times 3}$ denotes a 3×3 identity matrix and σ_e^2 is a scaling factor. Substituting this expression in (4.24), we can see that maximizing the output SNR in (4.23) is equivalent to maximizing the power of $z(n)$ under the constraint

$$\mathbf{w}^\mathrm{T}\mathbf{w} = 1. \tag{4.26}$$

Therefore, our objective can be formulated as

$$\hat{\mathbf{w}} = \arg\max_\mathbf{w} \mathbf{w}^\mathrm{T}\mathbf{\Phi}_\mathbf{v}\mathbf{w} \;\text{ s.t. }\; \mathbf{w}^\mathrm{T}\mathbf{w} = 1. \tag{4.27}$$

The optimal solution $\mathbf{w}_\mathrm{o}$ is given by $\mathbf{s}/\|\mathbf{s}\|_2$, where $\|\cdot\|_2$ denotes the ℓ-2 vector norm.

For the more general problem where the noise is not spherically white, the objective

function would be given by

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \mathbf{w}^{\mathrm{T}}\mathbf{\Phi}_{\mathbf{v}}\mathbf{w} \;\text{ s.t. }\; \mathbf{w}^{\mathrm{T}}\mathbf{\Phi}_{\mathbf{e}}\mathbf{w} = 1. \tag{4.28}$$

In this case an estimate of $\mathbf{\Phi}_{\mathbf{e}}$ would be required.

4.2.2 Eigenbeam-based particle velocity vector estimation

The noisy particle velocity vector $\mathbf{v}$ can be measured using an acoustic vector sensor (e.g., the Microflown [28]), however here we wish to measure it using conventional pressure sensors. We follow the approach presented in Section 4.1, i.e., the vector $\mathbf{v}(k)$ is computed using (4.15) and (4.16). The particle velocities $v_a(n)$ in the discrete time domain are then obtained by taking the inverse discrete Fourier transform of the signals $V_a(k)$ evaluated at discrete values of wavenumber k.

4.2.3 Adaptive localization algorithm

4.2.3.1 Gradient ascent algorithm for spherically white noise

The constraint optimization problem in (4.27) can be solved using the method of Lagrange multipliers:

$$\mathcal{L}(\mathbf{w}, \lambda) = \mathbf{w}^{\mathrm{T}}\mathbf{\Phi}_{\mathbf{v}}\mathbf{w} + \lambda\left(\mathbf{w}^{\mathrm{T}}\mathbf{w} - 1\right), \tag{4.29}$$

where $\mathcal{L}$ denotes the Lagrangian and λ denotes the Lagrange multiplier. The update equation is given by

$$\hat{\mathbf{w}}(n) = \hat{\mathbf{w}}(n-1) + \mu\nabla L_{\mathbf{w}}|_{\mathbf{w}=\hat{\mathbf{w}}(n-1)}, \tag{4.30}$$

where μ is the step size and

$$\nabla L_{\mathbf{w}} = 2\mathbf{\Phi}_{\mathbf{v}}\mathbf{w} + \lambda\mathbf{w}. \tag{4.31}$$

We determine λ under the constraint that $\mathbf{w}^{\mathrm{T}}(n)\mathbf{w}(n) = 1$, neglecting terms of $\mathcal{O}(\mu^2)$,

as follows:

$$\begin{aligned}
\left[\mathbf{w}(n-1)+\mu\nabla L_{\mathbf{w}}\right]^{\mathrm{T}}\left[\mathbf{w}(n-1)+\mu\nabla L_{\mathbf{w}}\right] &= 1\\
\mathbf{w}^{\mathrm{T}}(n-1)\mathbf{w}(n-1)+\ 2\mu\mathbf{w}^{\mathrm{T}}(n-1)\left[2\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{w}(n-1)+\lambda\mathbf{w}(n-1)\right] &= 1\\
\mathbf{w}^{\mathrm{T}}(n-1)\mathbf{w}(n-1)+\ 4\mu\mathbf{w}^{\mathrm{T}}(n-1)\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{w}(n-1)+2\mu\lambda\mathbf{w}^{\mathrm{T}}(n-1)\mathbf{w}(n-1) &= 1\\
-2\mathbf{w}^{\mathrm{T}}(n-1)\boldsymbol{\Phi}_{\mathbf{v}}\mathbf{w}(n-1) &= \lambda,
\end{aligned}$$

where it has been assumed that $\mathbf{w}^{\mathrm{T}}(n-1)\mathbf{w}(n-1) = 1$. Now we obtain the update equation by substituting λ into (4.30)

$$\hat{\mathbf{w}}(n) = \hat{\mathbf{w}}(n-1)+\mu\left[2\boldsymbol{\Phi}_{\mathbf{v}}\hat{\mathbf{w}}(n-1)-2\hat{\mathbf{w}}^{\mathrm{T}}(n-1)\boldsymbol{\Phi}_{\mathbf{v}}\hat{\mathbf{w}}(n-1)\hat{\mathbf{w}}(n-1)\right]. \tag{4.32}$$

4.2.3.2 Sign ambiguity

PCA and the method described in Section 4.2.3.1 have an inherent sign ambiguity which is not mathematically solvable. To obtain an estimate $\hat{\mathbf{u}}$ of $\mathbf{u}$ that points in the correct direction, we need to determine the correct sign from an analysis of the data. This can be done by looking at the sign of the correlation r_{zp} between $z(n)$ and $p(n)$: if it is positive, then $\mathbf{u}$ points in the opposite direction to $\mathbf{w}$, and if it is negative, then $\mathbf{u}$ points in the same direction as $\mathbf{w}$:

$$\mathbf{u}(n) = -\mathrm{sign}\left\{r_{zp}\right\}\mathbf{w}(n). \tag{4.33}$$

4.2.3.3 Implementation

For an efficient implementation which allows for tracking, we do not perform the processing on a per sample basis, but instead on a frame by frame basis, where ℓ denotes the frame index. We initialize the algorithm for frame $\ell = 0$ using a standard PCA, i.e., we take the eigenvector corresponding to the largest eigenvalue of the data covariance matrix $\boldsymbol{\Phi}_{\mathbf{v}}(0)$.

Let τ_{l} denote the frame length and τ_{inc} the frame increment, thus yielding an overlap

of 75% for $\tau_{\text{inc}} = \tau_1/4$ for example. The covariance matrix $\mathbf{\Phi}_{\mathbf{v}}(\ell)$ can be recursively estimated over τ_1 samples using

$$\hat{\mathbf{\Phi}}_{\mathbf{v}}(\ell) = \beta_v \hat{\mathbf{\Phi}}_{\mathbf{v}}(\ell-1) + (1-\beta_v)\frac{1}{\tau_1}\sum_{n=\ell\tau_{\text{inc}}}^{\ell\tau_{\text{inc}}+\tau_1-1}\mathbf{v}(n)\mathbf{v}^{\mathrm{T}}(n), \tag{4.34}$$

where β_v is a weighting factor: the larger the weighting factor, the larger the contribution of previous samples. The correlation $r_{zp}(\ell)$ can similarly be estimated as

$$\hat{r}_{zp}(\ell) = \beta_{zp}\,\hat{r}_{zp}(\ell-1) + (1-\beta_{zp})\frac{1}{\tau_1}\sum_{n=\ell\tau_{\text{inc}}}^{\ell\tau_{\text{inc}}+\tau_1-1}\hat{\mathbf{w}}^{\mathrm{T}}(n)\mathbf{v}(n)\,p(n), \tag{4.35}$$

where β_{zp} is a weighting factor similar to β_v.

The update equation for $\hat{\mathbf{w}}$ is given by:

$$\hat{\mathbf{w}}(\ell) = \hat{\mathbf{w}}(\ell-1) + 2\mu\left[\hat{\mathbf{\Phi}}_{\mathbf{v}}(\ell)\hat{\mathbf{w}}(\ell-1) - \hat{\mathbf{w}}^{\mathrm{T}}(\ell-1)\hat{\mathbf{\Phi}}_{\mathbf{v}}(\ell)\hat{\mathbf{w}}(\ell-1)\hat{\mathbf{w}}(\ell-1)\right]. \tag{4.36}$$

Finally the estimated unit vector pointing from the sensor towards the source, for frame ℓ, is given by:

$$\hat{\mathbf{u}}(\ell) = -\text{sign}\left\{\hat{r}_{zp}(\ell)\right\}\hat{\mathbf{w}}(\ell). \tag{4.37}$$

4.2.4 Performance evaluation

4.2.4.1 Experiment setup

We tested our algorithm in a room acoustics scenario simulated using SMIRgen [54], an implementation of the AIR simulation algorithm presented in Chapter 3. The receiver, a 32 microphone rigid spherical microphone array of radius 4.2 cm (the same specifications as the *Eigenmike*), was placed near the centre of a $4 \times 6 \times 8$ m room. We limited the AIRs to 2048 samples, with a sampling frequency of 8 kHz. The source signal consisted of 2 s of white Gaussian noise. In order to model sensor noise, spatio-temporally white Gaussian noise was added to the microphone signals; the noise power was set such that

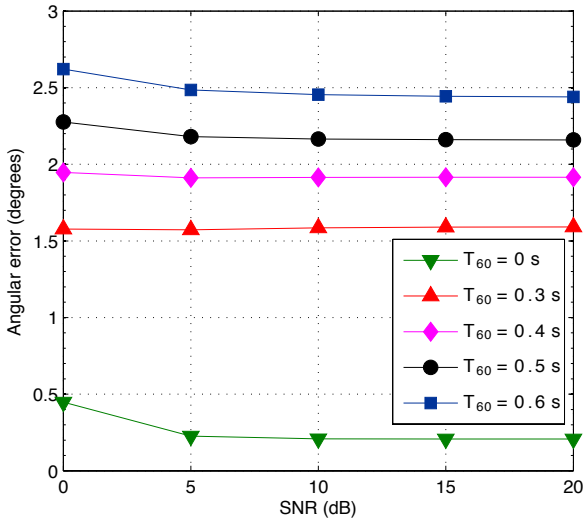


Figure 4.6: Position error as a function of SNR and reverberation time T_{60}, for the first experiment where the source was static. © 2010 IEEE.

a given SNR was obtained at the microphone closest to the source.

4.2.4.2 Static source

In a first experiment for a static source, we performed Monte Carlo simulations with 10 runs, for various SNRs and room reverberation times T_{60}. For each run a new source position was randomly selected, at a distance of 1.5 m from the centre of the array. We chose a step size $\mu = 1$, weighting factors $\beta_v = 0.95$ and $\beta_{zp} = 0.98$, frame length $\tau_1 = 256$ and frame increment $\tau_{\text{inc}} = 64$.

To evaluate the performance of our algorithm we computed the angular error ϵ, which is the angle between a unit vector $\mathbf{u}$ pointing in the correct direction and a unit vector $\hat{\mathbf{u}}$ pointing in the estimated direction, using (4.19).

The angular error averaged over all estimates from 1.5 to 2 s is shown in Fig. 4.6. It can be seen that even with reverberation times up to 600 ms and SNRs as low as 0 dB, the angular error remains below 3.5°. It should be noted that the error is larger than in Section 4.1, as we have sacrificed some accuracy for improved time resolution, which is a reasonable tradeoff in a tracking application.

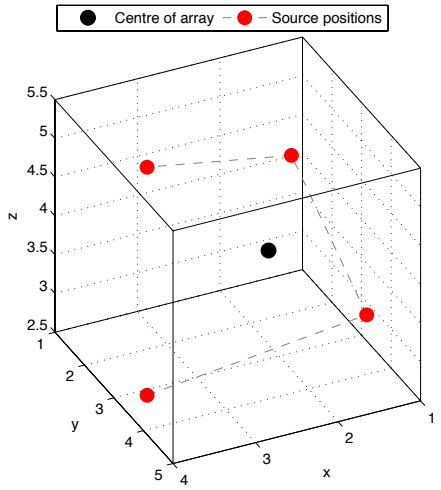


Figure 4.7: Source positions relative to centre of spherical microphone array, for the second experiment where the source was moving. © 2010 IEEE.

4.2.4.3 Moving source

In a second experiment, over a period of 2 s we placed the source in four different positions around the array, at a distance of 1.5–2 m from the centre of the array, as illustrated in Fig. 4.7. We chose $\mu = 0.3$, $\beta_v = 0.9$, $\beta_{zp} = 0.95$, $\tau_1 = 128$ and $\tau_{\text{inc}} = 32$.

The reference and estimated source positions are shown in Fig. 4.8 for various reverberation times and an SNR of 5 dB. After an initial tracking time, the estimates converge to the true position, within a couple of degrees. The results are similar for SNRs above 5 dB. While the tracking time generally increases as the reverberation time increases, after tracking the accuracy of the estimates is good even for high reverberation times. It should be noted that while in some cases it appears the estimate is diverging from the true position (e.g., for the azimuth at 500–700 ms), this is due to the sign ambiguity: once the sign has changed (e.g., at 600 ms), it can be seen that the estimate is actually converging towards the true position.

4.2.4.4 Choice of adaptive parameters

If we wish to reduce the tracking time, we can increase μ and decrease β_v and β_{zp}, at the risk of creating instability and at the expense of accuracy. If we wish to increase the

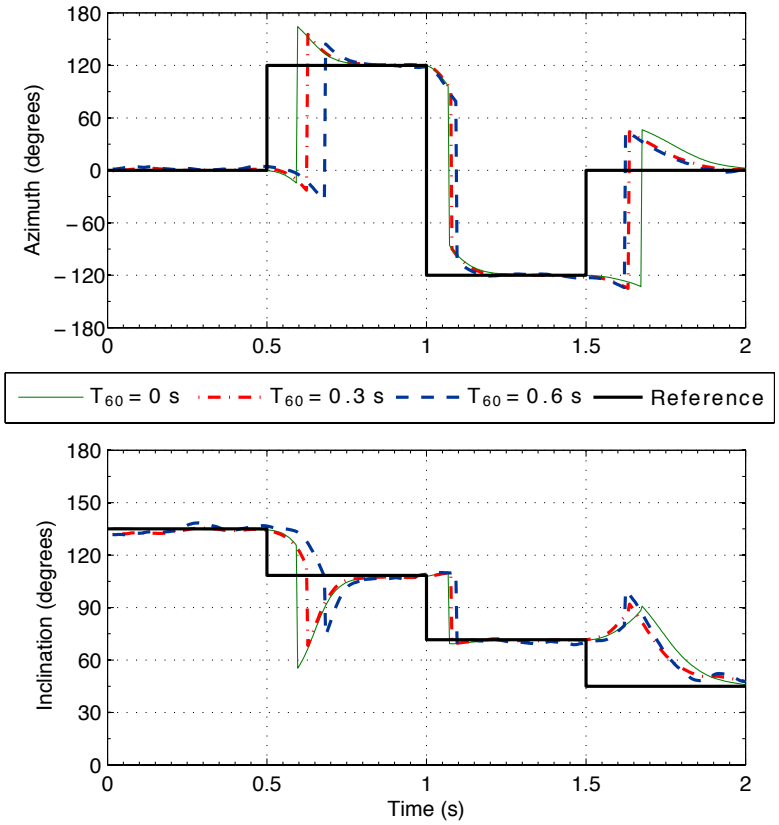


Figure 4.8: Reference and estimated source positions as a function of time, for various reverberation times T_{60}, for the second experiment where the source was moving. © 2010 IEEE.

accuracy, we can increase β_v and β_{zp} and decrease μ, at the expense of a longer tracking time.

4.2.5 Conclusion

The proposed algorithm allows us to track sources in two dimensions (azimuth and inclination) using a spherical microphone array. An evaluation of this algorithm has shown that it has high accuracy for the source-array distances considered, with angular errors of 1–3° after convergence, even in the presence of high levels of noise (down to SNRs of 0–5 dB), and reverberation times up to 600 ms.

4.3 Diffuseness estimation

The estimation of the diffuseness of a sound field is useful for a number of acoustic signal processing applications. For example, this information can be used in dereverberation algorithms to suppress diffuse reverberant energy while retaining the direct sound [51]. It can also be used to improve the accuracy of source localization algorithms, by eliminating inaccurate DOA estimates obtained under highly diffuse conditions. Moreover, the diffuseness represents an important parameter in the description of spatial sound, e.g., in Directional Audio Coding (DirAC) [87].

Diffuseness estimation has previously been accomplished by considering the spatial coherence between a pair of omnidirectional microphones [115] and an arbitrary pair of first-order microphones [114]. Spherical microphone arrays, typically incorporating a few dozen microphones, enable the analysis of sound fields in three dimensions [1, 77], and have recently been used for speech enhancement applications such as noise reduction [56] and dereverberation [60].

In this section, we take advantage of the availability of these additional microphone signals, and propose a diffuseness estimation algorithm for spherical microphone arrays based on the coherence between eigenbeams. An earlier version of this work was previously published in [64].

In the spatial domain, the omnidirectional microphone signals are correlated at low frequencies even when the sound field is purely diffuse, which makes robust diffuseness estimation difficult. An advantage of the SHD is that in a purely diffuse sound field, the coherence between the eigenbeams is zero, while in a purely directional sound field (i.e., due to a single plane wave) the coherence is one. We also take advantage of the availability of many eigenbeam pairs to reduce the variance of our estimates.

Portions of this work were first published in the *Proceedings of the IEEE Convention of Electrical and Electronics Engineers in Israel (IEEEI)* [64] in 2012. © 2012 IEEE.

4.3.1 Problem formulation

In the following, we work in the STFT domain, where k denotes the discrete frequency index and ℓ denotes the time frame index.

4.3.1.1 Spatial and spherical harmonic domain signal models

In the spatial domain, the signal $X(k, \mathbf{r}, \ell)$ received at a microphone position $\mathbf{r} = (r, \Omega) = (r, \theta, \phi)$ (in spherical coordinates) is modeled as the sum of a *directional signal* X_{dir}, a *diffuse signal* X_{diff} and a *sensor noise signal* V, i.e.,

$$X(k, \mathbf{r}, \ell) = X_{\text{dir}}(k, \mathbf{r}, \ell, \Omega_{\text{dir}}) + X_{\text{diff}}(k, \mathbf{r}, \ell) + V(k, \mathbf{r}, \ell). \tag{4.38}$$

The directional signal X_{dir} corresponds to a plane wave incident from a DOA Ω_{dir}. The diffuse signal X_{diff} is composed of an infinite number of independent plane waves with equal amplitude, random phase and uniformly distributed DOA [69]. The powers of the directional and diffuse signals received at a (virtual) omnidirectional reference microphone $\mathcal{M}_{\text{ref}}$ placed at the centre of the array are denoted as $P_{\text{dir}}(k, \ell)$ and $P_{\text{diff}}(k, \ell)$, respectively.

When dealing with spherical microphone arrays, it is convenient to work in the SHD, particularly for rigid arrays whose scattering behaviour can be described analytically in the SHD. We denote the spherical Fourier transform of $X(k, \mathbf{r}, \ell)$, as defined in (2.4), as $X_{lm}(k, \ell)$. In the following, we assume perfect spatial sampling; the effects of spatial aliasing [94] are therefore neglected.

Using the spherical Fourier transform in (2.4), the spatial domain signal model (4.38) can now be expressed in the SHD:

$$X_{lm}(k, \ell) = X_{lm}^{\text{dir}}(k, \ell, \Omega_{\text{dir}}) + X_{lm}^{\text{diff}}(k, \ell) + V_{lm}(k, \ell), \tag{4.39}$$

where X_{lm}, X_{lm}^{dir}, X_{lm}^{diff} and V_{lm} are respectively the spherical Fourier transforms of X,

X_{dir}, X_{diff} and V.

The directional signal $X_{lm}^{\text{dir}}(\mathbb{k}, \ell)$ can be expressed in the SHD as [90]

$$X_{lm}^{\text{dir}}(\mathbb{k}, \ell, \Omega_{\text{dir}}) = \sqrt{P_{\text{dir}}(\mathbb{k}, \ell)}\varphi_{\text{dir}}(\mathbb{k}, \ell)4\pi B_l(\mathbb{k})Y_{lm}^*(\Omega_{\text{dir}}), \tag{4.40}$$

where $\varphi_{\text{dir}}(\mathbb{k}, \ell)$ is the wave phase (with $|\varphi_{\text{dir}}(\mathbb{k}, \ell)| = 1\ \forall \mathbb{k}, \ell$). The mode strength $B_l(\mathbb{k})$ is given by evaluating the mode strength $b_l(k)$, as defined in Section 2.4, at discrete wavenumber values corresponding to the frequency indices $\mathbb{k}$. It is a function of the array properties (configuration, microphone type, radius); mode strength expressions for various configurations (open, rigid, dual-sphere, etc.) can be found in [92][4].

The diffuse signal $X_{lm}^{\text{diff}}(\mathbb{k}, \ell)$ is expressed in the SHD as

$$X_{lm}^{\text{diff}}(\mathbb{k}, \ell) = \sqrt{\frac{P_{\text{diff}}(\mathbb{k}, \ell)}{4\pi}} \int_{\Omega \in \mathcal{S}^2} \varphi_{\text{diff}}(\mathbb{k}, \ell, \Omega)4\pi B_l(\mathbb{k})Y_{lm}^*(\Omega)\mathrm{d}\Omega, \tag{4.41}$$

where $\varphi_{\text{diff}}(\mathbb{k}, \ell, \Omega)$ is the phase of the wave with DOA Ω (with $|\varphi_{\text{diff}}(\mathbb{k}, \ell, \Omega)| = 1\ \forall \mathbb{k}, \ell, \Omega$). As the plane waves are independent, the wave phases satisfy the property

$$\mathrm{E}\left[\varphi_{\text{diff}}(\mathbb{k}, \ell, \Omega)\varphi_{\text{diff}}^*(\mathbb{k}, \ell, \Omega')\right] = \delta_{\Omega - \Omega'}, \tag{4.42}$$

where δ is the Kronecker delta and $\mathrm{E}[\cdot]$ denotes mathematical expectation.

The signal received at the reference microphone $\mathcal{M}_{\text{ref}}$ is given by $X_{00}(\mathbb{k}, \ell)/\left[\sqrt{4\pi}B_0(\mathbb{k})\right]$ [55] (see Appendix C for derivation). Using this relationship and the fact that $|Y_{00}(\cdot)|^2 = (4\pi)^{-1}$, it can be verified that the powers of the directional and diffuse signals received at $\mathcal{M}_{\text{ref}}$ are given by P_{dir} and P_{diff}, respectively.

[4] It should be noted that in (4.40) and the expressions that follow, we have extracted the 4π scaling factor from the mode strength given in [92].

4.3.1.2 Signal to diffuse ratio and diffuseness

The signal-to-diffuse ratio (SDR) Γ at $\mathcal{M}_{\text{ref}}$ is given by

$$\Gamma(\mathbb{k}, \ell) = \frac{|X_{00}^{\text{dir}}(\mathbb{k}, \ell, \Omega_{\text{dir}})|^2}{\mathrm{E}\left[|X_{00}^{\text{diff}}(\mathbb{k}, \ell)|^2\right]} = \frac{P_{\text{dir}}(\mathbb{k}, \ell)}{P_{\text{diff}}(\mathbb{k}, \ell)}. \tag{4.43}$$

The *diffuseness* Ψ of the sound field can be defined as [29]

$$\Psi(\mathbb{k}, \ell) = \left[1 + \Gamma(\mathbb{k}, \ell)\right]^{-1}. \tag{4.44}$$

We have $\Psi(\mathbb{k}, \ell) \in [0, 1]$, where a diffuseness of 0 is obtained for $\Gamma(\mathbb{k}, \ell) \to \infty$ (purely directional field), 1 for $\Gamma(\mathbb{k}, \ell) = 0$ (purely diffuse field), and 0.5 for $\Gamma(\mathbb{k}, \ell) = 1$ (equal energy directional and diffuse fields).

In the following, we aim to estimate the diffuseness in (4.44) from the sound field observed using a spherical array.

4.3.2 Signal-to-diffuse ratio estimation using spatial coherence

In this section, we propose a method to estimate the SDR using the spatial coherence between the SHD signals (i.e., the eigenbeams). The estimated SDRs are then mapped to obtain the estimated diffuseness values using (4.44).

4.3.2.1 Spatial coherence

The complex spatial coherence between the eigenbeams $X_{lm}(\mathbb{k}, \ell)$ and $X_{l'm'}(\mathbb{k}, \ell)$ is defined for $(l, m) \neq (l', m')$ as

$$\gamma_{lm,l'm'}(\mathbb{k}, \ell) = \frac{\Phi_{lm,l'm'}(\mathbb{k}, \ell)}{\sqrt{\Phi_{lm,lm}(\mathbb{k}, \ell)}\sqrt{\Phi_{l'm',l'm'}(\mathbb{k}, \ell)}}, \tag{4.45}$$

where the power spectral densities (PSDs) Φ are given by

$$\Phi_{lm,l'm'}(\mathbb{k}, \ell) = \mathrm{E}\left[X_{lm}(\mathbb{k}, \ell) X_{l'm'}^{*}(\mathbb{k}, \ell)\right]. \tag{4.46}$$

We now determine expressions for the spatial coherence in purely directional and purely diffuse fields, in order to express the coherence in a *mixed* field as a function of the SDR Γ.

For purely directional sound, using (4.40) and (4.46) the PSD $\Phi^{\text{dir}}_{lm,l'm'}$ is expressed as

$$\Phi^{\text{dir}}_{lm,l'm'}(\Bbbk, \ell) = P_{\text{dir}}(\Bbbk, \ell)(4\pi)^2 B_l(\Bbbk) B^*_{l'}(\Bbbk) Y^*_{lm}(\Omega_{\text{dir}}) Y_{l'm'}(\Omega_{\text{dir}}) \tag{4.47}$$

and the directional field coherence $\gamma^{\text{dir}}_{lm,l'm'}$ is given by

$$\gamma^{\text{dir}}_{lm,l'm'}(\Bbbk, \ell) = \frac{B_l(\Bbbk) B^*_{l'}(\Bbbk) Y^*_{lm}(\Omega_{\text{dir}}) Y_{l'm'}(\Omega_{\text{dir}})}{\left| B_l(\Bbbk) B^*_{l'}(\Bbbk) Y^*_{lm}(\Omega_{\text{dir}}) Y_{l'm'}(\Omega_{\text{dir}}) \right|}. \tag{4.48}$$

For purely directional sound, the coherence $\gamma^{\text{dir}}_{lm,l'm'}$ therefore has unit magnitude.

For purely diffuse sound, using (4.41), (4.46) and the orthonormality of the spherical harmonics in (2.6), the PSD $\Phi^{\text{diff}}_{lm,l'm'}$ is expressed as

$$\begin{aligned} \Phi^{\text{diff}}_{lm,l'm'}(\Bbbk, \ell) &= P_{\text{diff}}(\Bbbk, \ell) 4\pi \int_{\Omega \in S^2} B_l(\Bbbk) B^*_{l'}(\Bbbk) Y^*_{lm}(\Omega) Y_{l'm'}(\Omega) \mathrm{d}\Omega \\ &= P_{\text{diff}}(\Bbbk, \ell) 4\pi B_l(\Bbbk) B^*_{l'}(\Bbbk) \delta_{l-l'} \delta_{m-m'}. \end{aligned} \tag{4.49a}$$

The diffuse field coherence $\gamma^{\text{diff}}_{lm,l'm'}$ is then given by

$$\gamma^{\text{diff}}_{lm,l'm'}(\Bbbk, \ell) = \frac{B_l(\Bbbk) B^*_{l'}(\Bbbk)}{|B_l(\Bbbk)||B_{l'}(\Bbbk)|} \delta_{l-l'} \delta_{m-m'} = 0, \tag{4.50}$$

providing $(l, m) \neq (l', m')$.

The sensor noise V is assumed to be spatially incoherent noise of equal power P^{N} at each of the Q equidistant microphones. The SHD noise V_{lm} is therefore also incoherent

across l and m and the PSD $\Phi^{\mathrm{N}}_{lm,l'm'}$ of the noise can be expressed as [119, eqn. 7.31]

$$\Phi^{\mathrm{N}}_{lm,l'm'}(\mathbb{k},\ell) = \mathrm{E}\left[V_{lm}(\mathbb{k},\ell)\,V^{*}_{l'm'}(\mathbb{k},\ell)\right] \tag{4.51a}$$

$$= P^{\mathrm{N}}\frac{4\pi}{Q}\delta_{l-l'}\delta_{m-m'}. \tag{4.51b}$$

The power of the noise at the reference microphone $\mathcal{M}_{\mathrm{ref}}$ is then given by $P^{\mathrm{N}}/\left[Q\,|B_0(\mathbb{k})|^2\right]$, i.e., it has been reduced by a factor $Q\,|B_0(\mathbb{k})|^2$.

In a mixed sound field, both the directional and diffuse sound fields X_{dir} and X_{diff} are present, in addition to incoherent noise V. We assume they are mutually uncorrelated, such that the PSD $\Phi_{lm,l'm'}$ is equal to the sum of the individual PSDs, i.e.,

$$\Phi_{lm,l'm'}(\mathbb{k},\ell) = \Phi^{\mathrm{dir}}_{lm,l'm'}(\mathbb{k},\ell) + \Phi^{\mathrm{diff}}_{lm,l'm'}(\mathbb{k},\ell) + \Phi^{\mathrm{N}}_{lm,l'm'}(\mathbb{k},\ell). \tag{4.52}$$

We define the *noiseless* coherence as

$$\gamma'_{lm,l'm'}(\mathbb{k},\ell) = \frac{\Phi'_{lm,l'm'}(\mathbb{k},\ell)}{\sqrt{\Phi'_{lm,lm}(\mathbb{k},\ell)}\sqrt{\Phi'_{l'm',l'm'}(\mathbb{k},\ell)}}, \tag{4.53}$$

where the noiseless PSD $\Phi'_{lm,l'm'}(\mathbb{k},\ell)$ is defined as $\Phi'_{lm,l'm'}(\mathbb{k},\ell) = \Phi^{\mathrm{dir}}_{lm,l'm'}(\mathbb{k},\ell) + \Phi^{\mathrm{diff}}_{lm,l'm'}(\mathbb{k},\ell)$. Using (4.47) and (4.49), the noiseless PSD can be expressed as

$$\Phi'_{lm,l'm'}(\mathbb{k},\ell) = 4\pi B_l(\mathbb{k})B^{*}_{l'}(\mathbb{k})\left[4\pi P_{\mathrm{dir}}(\mathbb{k},\ell)\,Y^{*}_{lm}(\Omega_{\mathrm{dir}})\,Y_{l'm'}(\Omega_{\mathrm{dir}}) + P_{\mathrm{diff}}(\mathbb{k},\ell)\delta_{l-l'}\delta_{m-m'}\right]. \tag{4.54}$$

By substituting (4.54) in (4.53), and using (4.43), it can straightforwardly be shown that

$$\gamma'_{lm,l'm'}(\mathbb{k},\ell) = \frac{\Gamma(\mathbb{k},\ell)\gamma^{\mathrm{dir}}_{lm,l'm'}(\mathbb{k},\ell)c_{lm}c_{l'm'}}{\sqrt{\Gamma^2(\mathbb{k},\ell)c^2_{lm}c^2_{l'm'} + \Gamma(\mathbb{k},\ell)(c^2_{lm}+c^2_{l'm'}) + 1}}, \tag{4.55}$$

where we have defined $c_{lm} = \sqrt{4\pi}|Y_{lm}(\Omega_{\mathrm{dir}})|$.

The noiseless PSDs in (4.53) cannot be directly observed, however as the noise V_{lm}

is incoherent across l and m, with sufficient time averaging the noise cross PSD $\Phi^{\mathrm{N}}_{lm,l'm'}$ will average to zero in $\Phi_{lm,l'm'}$. The noiseless auto PSD can be estimated providing an estimate of the noise power P^{N} is available. For simplicity, in this work we will assume a sufficiently high SNR and estimate the noiseless coherence directly from the noisy signals, i.e., we will not compensate for the noise. The effect of sensor noise on the estimation will be discussed in Section 4.3.4.

4.3.2.2 Signal-to-diffuse ratio estimation

The SDR is determined by first computing the coherence between pairs of eigenbeams $X_{lm}(\mathfrak{k},\ell)$ and $X_{l'm'}(\mathfrak{k},\ell)$. The SDR for each specific eigenbeam pair is then found by solving for $\Gamma(\mathfrak{k},\ell)$ in (4.55)[5], as in [114]:

$$\hat{\Gamma}_{lm,l'm'} = \frac{G + \sqrt{G^2 + 4\left(\left|\gamma'_{lm,l'm'}\right|^{-2} - 1\right)}}{2c_{lm}c_{l'm'}\left(\left|\gamma'_{lm,l'm'}\right|^{-2} - 1\right)}, \tag{4.56}$$

where we have defined

$$G = \frac{c_{lm}}{c_{l'm'}} + \frac{c_{l'm'}}{c_{lm}}. \tag{4.57}$$

In order to compute c_{lm}, the DOA Ω_{dir} must be estimated; a robust DOA estimation method for spherical arrays is presented in Section 4.1.

The possible combinations of the pair (l,m) form a set $\mathcal{A}$ with $(L+1)^2$ elements, where L is the array order. The SDR can be estimated using (4.56) for all possible combinations of (l,m) and (l',m') (i.e., the set $\mathcal{A}^2$) excluding identical pairs for which $(l,m) = (l',m')$; however we also exclude duplicate pairs $((l',m'),(l,m))$ that provide the same information as $((l,m),(l',m'))$ due to the symmetry of the coherence function. The reduced set thereby obtained is denoted as $\tilde{\mathcal{L}}$ and contains $\left[(L+1)^4 - (L+1)^2\right]/2$ elements.

[5]The dependencies on $\mathfrak{k}$ and ℓ have been omitted for brevity.

We then form an estimate of the SDR $\hat{\Gamma}$ by taking a weighted average of the SDR estimates $\hat{\Gamma}_{lm,l'm'}$, i.e.,

$$\hat{\Gamma}(\mathrm{k}, \ell) = \sum_{(l,m,l',m')\in\bar{\mathcal{L}}} \alpha_{lm,l'm'}(\mathrm{k})\hat{\Gamma}_{lm,l'm'}(\mathrm{k}, \ell), \tag{4.58}$$

where $\alpha_{lm,l'm'}$ is a normalized weighting function. Ideally, the optimal weights $\alpha^{\mathrm{opt}}_{lm,l'm'}$ depend on the variances of the SDR estimates. Since the variances are usually unknown, we propose to compute the weights as the geometric mean of the SNRs of the eigenbeams involved, i.e.,

$$\alpha_{lm,l'm'}(\mathrm{k}) = \frac{\sqrt{\mathrm{SNR}_{lm}(\mathrm{k})\mathrm{SNR}_{l'm'}(\mathrm{k})}}{\sum_{(l,m,l',m')\in\bar{\mathcal{L}}}\sqrt{\mathrm{SNR}_{lm}(\mathrm{k})\mathrm{SNR}_{l'm'}(\mathrm{k})}}, \tag{4.59}$$

where SNR_{lm} denotes the SNR at order l and degree m and is defined as

$$\mathrm{SNR}_{lm}(\mathrm{k}) = \frac{\left|X^{\mathrm{dir}}_{lm}(\mathrm{k}, \ell, \Omega_{\mathrm{dir}})\right|^2}{\mathrm{E}\left[\left|V_{lm}(\mathrm{k}, \ell)\right|^2\right]} \tag{4.60a}$$

$$= \left(P^{\mathrm{N}}\right)^{-1} 4\pi Q P_{\mathrm{dir}}(\mathrm{k}, \ell)\left|B_l(\mathrm{k})Y^*_{lm}(\Omega_{\mathrm{dir}})\right|^2. \tag{4.60b}$$

The weighting function can then be simplified to

$$\alpha_{lm,l'm'}(\mathrm{k}) = \frac{\left|B_l(\mathrm{k})B_{l'}(\mathrm{k})Y^*_{lm}(\Omega_{\mathrm{dir}})Y^*_{l'm'}(\Omega_{\mathrm{dir}})\right|}{\sum_{(l,m,l',m')\in\bar{\mathcal{L}}}\left|B_l(\mathrm{k})B_{l'}(\mathrm{k})Y^*_{lm}(\Omega_{\mathrm{dir}})Y^*_{l'm'}(\Omega_{\mathrm{dir}})\right|}. \tag{4.61}$$

Due to the chosen SNR definition, (4.61) depends only on the DOA and not on the wave or noise powers.

The weighted averaging of the SDR estimates, which is not performed in spatial domain coherence-based approaches with two microphones, aims to reduce the estimate variance, at the expense of increased computational complexity.

4.3.3 Diffuseness estimation using the pseudointensity vector

We compare the proposed (coherence-based) method with the previously proposed coefficient of variation (CV) method [4]. The CV method exploits the temporal variation of the intensity vector $\mathcal{I}$, and estimates the diffuseness as

$$\Psi_{\mathrm{CV}}(k,\ell) = \sqrt{1 - \frac{\|\mathrm{E}\left[\mathcal{I}(k,\ell)\right]\|_2}{\mathrm{E}\left[\|\mathcal{I}(k,\ell)\|_2\right]}}, \tag{4.62}$$

where $\|\cdot\|_2$ denotes the ℓ-2 vector norm.

As shown in Section 4.1, the intensity vector can be estimated using a linear combination of first-order eigenbeams obtained with a spherical microphone array. The resulting vector, which is proportional to the intensity vector, is called a*pseudointensity vector*. The reader is referred to Section 4.1 for details of the computation of the pseudointensity vector from X_{00}, $X_{1(-1)}$, X_{10} and X_{11}. We hereafter refer to the estimation of the diffuseness using the CV method based on pseudointensity vectors as the *modified* CV method.

It should be noted that while the modified CV method only makes use of first-order eigenbeams, all Q microphone signals are used to compute the pseudointensity vector, unlike in previous approaches where the intensity vector was estimated using either an acoustic vector sensor or four pressure microphones.

4.3.4 Performance evaluation

In this section, we evaluate the performance of the proposed SHD coherence-based method, and compare it to the performance of the modified CV method.

4.3.4.1 Experimental setup

We simulated the SHD signals received by a rigid spherical array of radius 4.2 cm up to an order L (either $L = 1$ or $L = 3$). The directional source signal consisted of complex white Gaussian noise, with a DOA of $(90°, 0°)$ (inclination, azimuth). This DOA was

assumed to be known for the estimation of the SDR in (4.56) and the weights in (4.61). The diffuse signal was generated by summing 200 plane waves with random phase and uniformly distributed DOAs; the diffuse signal power was set according to the desired SDR.

The noise signal consisted of additive complex white Gaussian noise; the noise power was set such that the desired SNR was obtained at the reference microphone $\mathcal{M}_{\text{ref}}$, i.e.,

$$\text{SNR} = \frac{\text{E}\left[|X_{00}^{\text{dir}}(k, \ell, \Omega_{\text{dir}})|^2\right]}{\text{E}\left[|V_{00}(k, \ell)|^2\right]}. \tag{4.63}$$

The noise power was therefore the same for all values of SDR. We chose to compute the SNR at $\mathcal{M}_{\text{ref}}$ because the directional signal power is different at each sensor, particularly for a rigid array. As noted in Section 4.3.2.1, the noise power at $\mathcal{M}_{\text{ref}}$ is reduced by a factor of $Q\,|B_0(k)|^2$ with respect to the sensors; the noise power at $\mathcal{M}_{\text{ref}}$ is therefore lowest at low frequencies, where $B_0(k)$ is highest. With $Q = 32$ microphones, at low frequencies an SNR of 25 dB at $\mathcal{M}_{\text{ref}}$ corresponds to an SNR of around 10 dB based on the noise power at the sensors.

Processing was performed in the STFT domain with a sampling frequency of 8 kHz, a window length of 16 ms and 50% overlap between consecutive frames, giving a hop length of $\tau_{\text{hop}} = 8$ ms. The expectations in (4.45) and (4.62) were estimated using moving averages over a given number of time frames N_{frames}, which is related to the time averaging length τ_{avg} via the expression $\tau_{\text{avg}} = (N_{\text{frames}} + 1)\,\tau_{\text{hop}}$. The performance results shown were averaged over 15 s of data.

4.3.4.2 Results

In Fig. 4.9 we plot the mean diffuseness estimated by the proposed and modified CV methods as a function of SDR, as well as the ideal diffuseness as given by (4.44). In this experiment, the time averaging length was 88 ms, and the proposed method exploited eigenbeams up to order $L = 3$. We find that for high SDRs, the proposed method

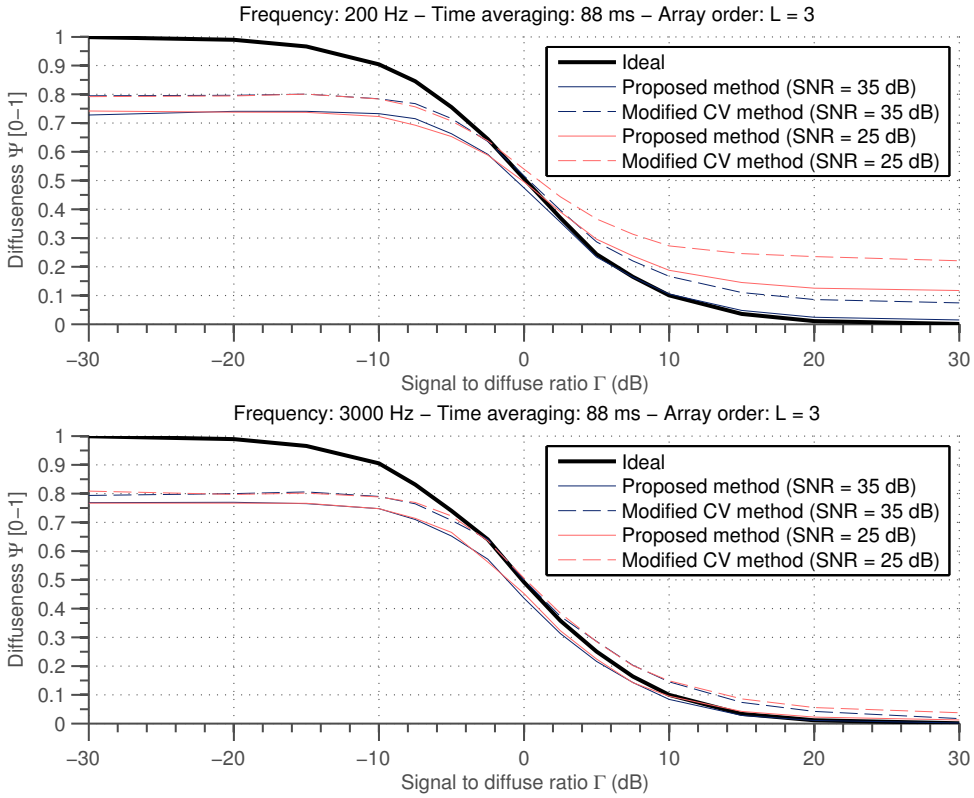


Figure 4.9: Mean diffuseness Ψ estimated using the proposed (coherence-based) method and the modified CV method, as a function of signal to diffuse ratio Γ, at two frequencies (200 Hz and 3 kHz) and two SNRs (25 dB and 35 dB). © 2012 IEEE.

estimates the diffuseness more accurately, particularly at low frequencies. For low SDRs, the proposed method has a slightly higher bias than the modified CV method, due to the limited time averaging, as in [114]. In addition, as the SNR decreases from 35 dB to 25 dB, for both methods the bias at low frequencies and high SDRs increases, however for the proposed method this bias is in part due to the lack of compensation for the noise power, as in [114].

We also plot the standard deviation of the diffuseness estimates as a function of the SDR in Fig. 4.10. It can be seen that at high SDRs, the estimates obtained using the proposed method have significantly lower variance than those obtained using the modified CV method, due to the averaging of the coherence estimates over all eigenbeam pairs. The proposed method's estimates also have lower variance at high frequencies and low SDRs.

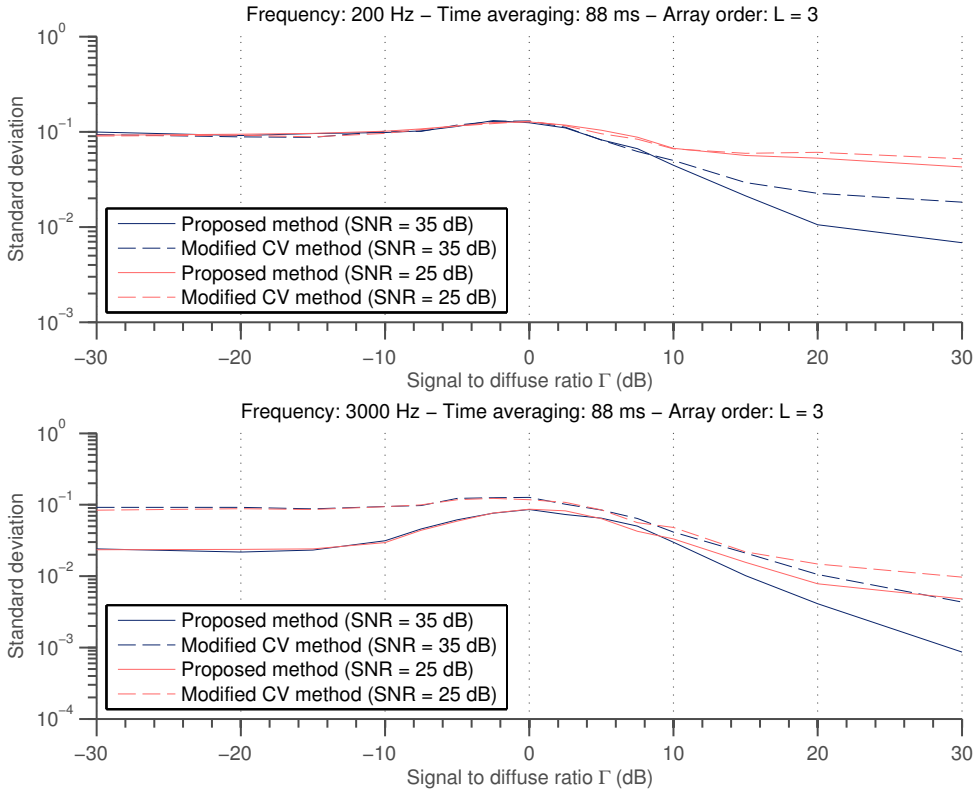


Figure 4.10: Standard deviation of the diffuseness estimates obtained using the proposed (coherence-based) method and the modified CV method, as a function of signal to diffuse ratio Γ, at two frequencies (200 Hz and 3 kHz) and two SNRs (25 dB and 35 dB). © 2012 IEEE.

In order to illustrate the effect of increasing the time averaging, in Fig. 4.11 we plot the mean diffuseness estimated by the two methods for two different averaging lengths (88 ms and 328 ms). As expected we see that the increase in time averaging significantly reduces the bias for the proposed method. With increased time averaging, the bias for the two methods is essentially the same at low SDRs, and is lower for the proposed method at high SDRs.

Finally in Fig. 4.12 we plot the standard deviation of the estimates obtained for array orders of $L = 1$ and $L = 3$. We find that by averaging over a larger number of SDR estimates, the variance of the final estimate is greatly reduced at low SDRs (except at low frequencies). We also note that even for $L = 1$, the variance of the proposed method's estimates is lower than those obtained using the modified CV method, which also uses

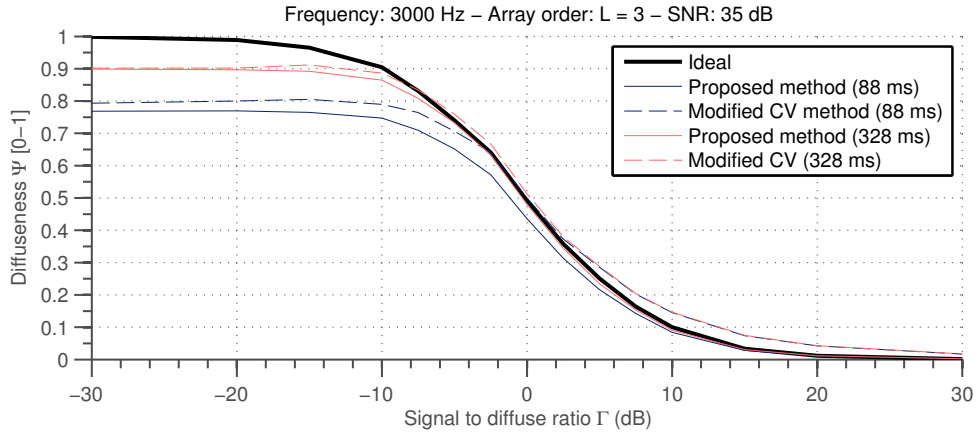


Figure 4.11: Mean diffuseness Ψ estimated using the proposed (coherence-based) method and the modified CV method, as a function of signal to diffuse ratio Γ, for two time averaging lengths (88 ms and 328 ms). © 2012 IEEE.

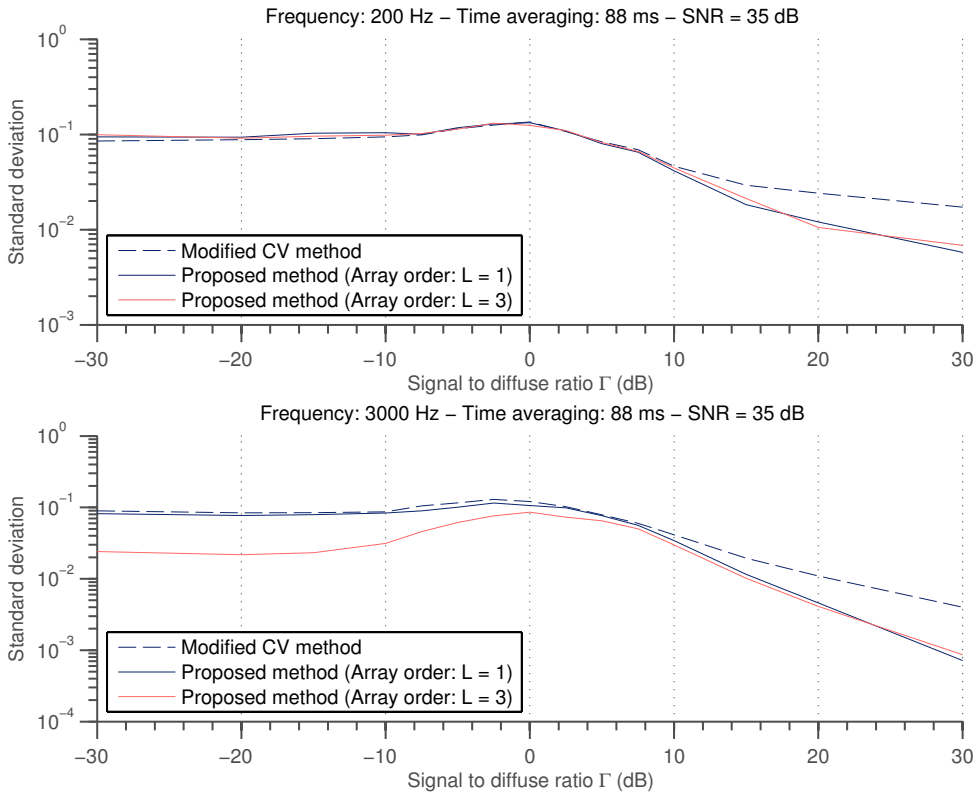


Figure 4.12: Standard deviation of the diffuseness estimates obtained using the proposed (coherence-based) method and the modified CV method, as a function of signal to diffuse ratio Γ, for two array orders ($L = 1$ and $L = 3$) at two frequencies (200 Hz and 3 kHz). © 2012 IEEE.

only zero- and first-order eigenbeams.

4.3.5 Conclusions

In this section, we proposed a diffuseness estimator based on the coherence between eigenbeams. We showed that at high SDRs, the proposed method has a lower bias than a previously proposed spatial domain method (the modified CV method), and that the underestimation of the diffuseness at low SDRs can be reduced by increased time averaging. Finally we found that increasing the array order significantly reduces the variance of the diffuseness estimates, and that even using a first-order array yields estimates with lower variance than those obtained with the modified CV method.

Chapter 5

Noise reduction

In many distant speech acquisition scenarios, such as hands-free telephony, hearing aids, or teleconferencing, the desired speech signal is corrupted by noise, such as sensor noise, diffuse noise or interfering speech. This noise can degrade both the speech quality and intelligibility, making communication difficult or even impossible. Noise reduction algorithms seek to mitigate these effects and extract the desired speech signal.

This objective is commonly achieved through the use of microphone arrays [11,17,36], which allow us to take advantage of the spatial properties of the sound field in order to achieve better noise reduction performance than with a single microphone. These microphone arrays are mostly two dimensional (planar). Spherical microphone arrays are advantageous due to their ability to analyze the sound field in three dimensions [1,35,90]; the captured sound field can then be efficiently described in the spherical harmonic domain (SHD), as presented in Chapter 2.

Over the past few decades, many spatio-temporal filters or *beamformers* have been proposed to process the signals received by microphone arrays in the spatial domain (see [11, 42, 50] and the references therein). SHD beamformers have more recently been proposed in which, instead of filtering and combining the individual microphone signals, we filter and sum the SHD signals (the *eigenbeams*) [3, 56, 109, 120].

Signal-dependent beamformers optimize the filter weights taking into account char-

acteristics of the speech and noise, as opposed to fixed beamformers, which apply a constraint to a specific look direction and optimize the filter weights with respect to performance measures such as white noise gain, sidelobe levels, or the directivity index. In this chapter, we propose a SHD tradeoff beamformer, which achieves a balance between noise reduction and speech distortion, controlled by a tradeoff parameter. For specific choices of this parameter, SHD equivalents of the well-known minimum variance distortionless response (MVDR) and multichannel Wiener filters are obtained.

In order to compute the weights of signal-dependent beamformers, we usually at least require an estimate of the noise power spectral density (PSD) matrix. Unfortunately, in practice the noise signals are not always observable and the noise PSD must be estimated from the noisy signals. Previously proposed spatial domain noise estimators based on the speech presence probability (SPP) [22, 47, 52, 105] seek to update the noise PSD estimate only in time-frequency bins where speech is absent. A recent contribution by Souden et al. [104] proposes a Gaussian model based multichannel SPP estimator, which is able to detect spatially coherent sound sources from any direction.

In this work, we seek to distinguish between desired coherent sources, which are sources located within a given region of interest, and undesired coherent sources (considered to be noise); however, this is not possible using only the SPP. To make this distinction, we need to take into account signal properties and/or spatial information. We propose to estimate the noise and desired PSD matrices using a desired speech presence probability (DSPP) estimator based on the product of the multichannel SPP [104] and a direction of arrival (DOA) dependent probability. The DOA-based probability is computed for each time-frequency bin by estimating the DOA using the pseudointensity vector method [57] (as presented in Sec. 4.1) and determining whether the active source is likely to lie within a desired range of DOAs, taking into account the variance of the DOA estimates. The desired range or ranges of DOAs are assumed to be known; they could be determined manually, or based on facial recognition and/or tracking data [117], for example. We then use the estimated PSD matrices to compute the weights of the

proposed tradeoff beamformer.

Earlier versions of this work were published in [56,59]. This work differs in a number of important ways: instead of using DOA estimates to control the *a priori* DSPP, the SPP is computed using a fixed *a priori* SPP and is then multiplied by a DOA-dependent probability to yield the DSPP; the uncertainty in the DOA estimates is taken into account; and the estimated statistics are applied to a tradeoff beamformer (which can be controlled by the DSPP) instead of an MVDR beamformer.

The remainder of this chapter is structured as follows: Sec. 5.1 describes the signal model and formulates the problem. Sec. 5.2 proposes a SHD tradeoff beamformer which is used to perform the noise reduction and depends on the signal statistics. Sec. 5.3 explains how the signal statistics can be estimated using the DSPP, Sec. 5.4 proposes a novel way of estimating the DSPP, and Sec. 5.5 summarizes the complete proposed statistics estimation algorithm. Sec. 5.6 evaluates the performance of the algorithm and of the tradeoff beamformer based on the estimated statistics. Finally, conclusions are provided in Sec. 5.7.

5.1 Signal model

5.1.1 Spatial domain signal model

We consider a scenario in which a spherical microphone array captures a mixture of desired speech originating from a source $\mathcal{S}$, spatially coherent noise (e.g., interfering speech), and background noise that can consist of a mixture of spatially incoherent noise (used to model sensor noise) and partially coherent noise (used to model spherically or cylindrically isotropic noise). Throughout this chapter, we work in the short-time Fourier transform (STFT) domain with a discrete frequency index k and a discrete time index ℓ[1].

[1]For brevity the time index is omitted in this section.

The spherical microphone array captures Q noisy signals $P(\mathbb{k}, \mathbf{r}_q)$ at microphone positions $\mathbf{r}_q = (r, \Omega_q)$ (in spherical coordinates), where r is the radius of the sphere and $q \in \{1, \ldots, Q\}$. The signal model is expressed as

$$\begin{aligned} P(\mathbb{k}, \mathbf{r}_q) &= H(\mathbb{k}, \mathbf{r}_q)S(\mathbb{k}) + V_{\mathrm{c}}(\mathbb{k}, \mathbf{r}_q) + V_{\mathrm{nc}}(\mathbb{k}, \mathbf{r}_q) \\ &= X(\mathbb{k}, \mathbf{r}_q) + V_{\mathrm{c}}(\mathbb{k}, \mathbf{r}_q) + V_{\mathrm{nc}}(\mathbb{k}, \mathbf{r}_q), \end{aligned} \tag{5.1}$$

where S is the source signal, X is the reverberant speech signal, V_{c} is the coherent noise signal, V_{nc} is the background noise signal, and $H(\mathbb{k}, \mathbf{r}_q)$ is the acoustic transfer function (ATF) between the source $\mathcal{S}$ and the microphone at angle Ω_q. The source $\mathcal{S}$ is located within a region of interest $\mathcal{R}$, while the coherent noise source(s) in V_c are located outside $\mathcal{R}$. The signals P, X, V_{c} and V_{nc} are a function of the microphone position, time and frequency, and are thus referred to as spatial domain signals; they are in addition also STFT domain signals.

The ATFs are assumed to be time-invariant. We also assume that the reverberant speech signals $X(\mathbb{k}, \mathbf{r}_q)$ and the noise signals $V_{\mathrm{c}}(\mathbb{k}, \mathbf{r}_q)$ and $V_{\mathrm{nc}}(\mathbb{k}, \mathbf{r}_q)$ are mutually uncorrelated. The reverberant speech signals $X(\mathbb{k}, \mathbf{r}_q)$ originate from a single source and are therefore, by definition, coherent at all microphones in the array.

5.1.2 Spherical harmonic domain signal model

When dealing with spherical microphone arrays, it is convenient to work in the spherical harmonic domain instead of the spatial domain. The spherical Fourier transform $F_{lm}(\mathbb{k})$ of a spatial domain signal $F(\mathbb{k}, \mathbf{r}_q)$ involves an integral over all angles Ω, however it can be approximated for a discretely sampled sound field using (2.4)

$$F_{lm}(\mathbb{k}) \approx \sum_{q=1}^{Q} c_q F(\mathbb{k}, \mathbf{r}_q) Y_{lm}^*(\Omega_q), \tag{5.2}$$

where Y_{lm} is the spherical harmonic of order $l \in \{0, \dots, L\}$ and degree $m \in \{-l, \dots, l\}$ and $(\cdot)^*$ denotes the complex conjugate. The weights c_q are chosen such that the approximation in (5.2) is as accurate as possible (c.f. [90] for examples); with a sufficient number of microphones and appropriate positioning, the error involved in this approximation can be eliminated entirely for a finite L. All spatial sampling schemes require at least $Q = (L+1)^2$ microphones to sample a sound field of order L without spatial aliasing. For more information on spatial sampling and aliasing, the reader is referred to Sec. 2.3.

We can now express our signal model in the SHD as:

$$\begin{aligned} P_{lm}(\mathfrak{k}) &= H_{lm}(\mathfrak{k})S(\mathfrak{k}) + V_{lm,\mathrm{c}}(\mathfrak{k}) + V_{lm,\mathrm{nc}}(\mathfrak{k}) \\ &= X_{lm}(\mathfrak{k}) + V_{lm,\mathrm{c}}(\mathfrak{k}) + V_{lm,\mathrm{nc}}(\mathfrak{k}), \end{aligned} \tag{5.3}$$

where $P_{lm}(\mathfrak{k})$, $H_{lm}(\mathfrak{k})$, $X_{lm}(\mathfrak{k})$, $V_{lm,\mathrm{c}}(\mathfrak{k})$ and $V_{lm,\mathrm{nc}}(\mathfrak{k})$ respectively denote the SHD representations of $P(\mathfrak{k}, \mathbf{r}_q)$, $H(\mathfrak{k}, \mathbf{r}_q)$, $X(\mathfrak{k}, \mathbf{r}_q)$, $V_{\mathrm{c}}(\mathfrak{k}, \mathbf{r}_q)$ and $V_{\mathrm{nc}}(\mathfrak{k}, \mathbf{r}_q)$.

5.1.3 Mode strength compensation

The eigenbeams P_{lm}, H_{lm}, X_{lm}, $V_{lm,\mathrm{c}}$ and $V_{lm,\mathrm{nc}}$ are dependent on the mode strength B_l, which is a function of the array properties (radius, configuration, microphone type) [92]. The mode strength $B_l(\mathfrak{k})$ is given by evaluating the mode strength $b_l(k)$, as defined in Section 2.4, at discrete wavenumber values corresponding to the frequency indices $\mathfrak{k}$. To cancel this dependence, the eigenbeams are divided by the mode strength to give mode strength compensated eigenbeams[2]:

$$\begin{aligned} \widetilde{P}_{lm}(\mathfrak{k}) &= \left[\sqrt{4\pi} B_l(\mathfrak{k})\right]^{-1} P_{lm}(\mathfrak{k}) \\ &= \widetilde{H}_{lm}(\mathfrak{k})S(\mathfrak{k}) + \widetilde{V}_{lm,\mathrm{c}}(\mathfrak{k}) + \widetilde{V}_{lm,\mathrm{nc}}(\mathfrak{k}) \\ &= \widetilde{X}_{lm}(\mathfrak{k}) + \widetilde{V}_{lm,\mathrm{c}}(\mathfrak{k}) + \widetilde{V}_{lm,\mathrm{nc}}(\mathfrak{k}), \end{aligned} \tag{5.4}$$

[2] It should be noted that in other parts of this thesis, the mode strength compensation was not explicitly performed, but was instead included in the beamformer weights, e.g., in (2.14).

where $\widetilde{P}_{lm}$, $\widetilde{H}_{lm}$, $\widetilde{X}_{lm}$, $\widetilde{V}_{lm,\mathrm{c}}$ and $\widetilde{V}_{lm,\mathrm{nc}}$ respectively denote the eigenbeams P_{lm}, H_{lm}, X_{lm}, $V_{lm,\mathrm{c}}$ and $V_{lm,\mathrm{nc}}$ after mode strength compensation.

With the addition of the $\sqrt{4\pi}$ scaling factor, $\widetilde{P}_{00}(\mathbb{k})$ is equal to the signal which would be received were an omnidirectional microphone $\mathcal{M}_{\mathrm{ref}}$ to be placed at a position corresponding to the centre of the sphere [55] (see Appendix C for derivation), i.e., at the origin of the spherical coordinate system. Our aim is to estimate the desired speech component $\widetilde{X}_{00}(\mathbb{k})$ of this signal using a tradeoff beamformer.

5.2 Tradeoff beamformer

In this section, we derive a signal-dependent tradeoff beamformer, which achieves a tradeoff between noise reduction and speech distortion. This tradeoff beamformer makes use of signal statistics that can be estimated using the method presented in the rest of this chapter.

It is convenient to rewrite the SHD signal model (5.4) in vector notation, where each of the vectors is of length $N = (L+1)^2$, the total number of eigenbeams up to order L:

$$\begin{aligned}\widetilde{\mathbf{p}}(\mathbb{k}) &= \widetilde{\mathbf{h}}(\mathbb{k})S(\mathbb{k}) + \widetilde{\mathbf{v}}_{\mathrm{c}}(\mathbb{k}) + \widetilde{\mathbf{v}}_{\mathrm{nc}}(\mathbb{k}) \\ &= \widetilde{\mathbf{x}}(\mathbb{k}) + \widetilde{\mathbf{v}}_{\mathrm{c}}(\mathbb{k}) + \widetilde{\mathbf{v}}_{\mathrm{nc}}(\mathbb{k}) \\ &= \mathbf{d}(\mathbb{k})\widetilde{X}_{00}(\mathbb{k}) + \widetilde{\mathbf{v}}(\mathbb{k}),\end{aligned} \tag{5.5}$$

where, as in the spatial domain [41, 50], $\mathbf{d}$ is a propagation vector of relative transfer functions given by

$$\mathbf{d}(\mathbb{k}) = \left[1 \; \frac{\widetilde{H}_{1(-1)}(\mathbb{k})}{\widetilde{H}_{00}(\mathbb{k})} \; \frac{\widetilde{H}_{10}(\mathbb{k})}{\widetilde{H}_{00}(\mathbb{k})} \; \frac{\widetilde{H}_{11}(\mathbb{k})}{\widetilde{H}_{00}(\mathbb{k})} \cdots \frac{\widetilde{H}_{LL}(\mathbb{k})}{\widetilde{H}_{00}(\mathbb{k})}\right]^{\mathrm{T}},$$

the vector $\widetilde{\mathbf{p}}$ is defined as

$$\widetilde{\mathbf{p}}(\mathbb{k}) = \left[\widetilde{P}_{00}(\mathbb{k})\, \widetilde{P}_{1(-1)}(\mathbb{k})\, \widetilde{P}_{10}(\mathbb{k})\, \widetilde{P}_{11}(\mathbb{k}) \cdots \widetilde{P}_{LL}(\mathbb{k})\right]^{\mathrm{T}},$$

$(\cdot)^{\mathrm{T}}$ denotes the vector transpose, and $\widetilde{\mathbf{x}}(\mathbb{k})$, $\widetilde{\mathbf{h}}(\mathbb{k})$, $\widetilde{\mathbf{v}}_{\mathrm{c}}(\mathbb{k})$ and $\widetilde{\mathbf{v}}_{\mathrm{nc}}(\mathbb{k})$ are defined similarly to $\widetilde{\mathbf{p}}(\mathbb{k})$. We assume $H_{00}(\mathbb{k}) \neq 0\ \forall \mathbb{k}$, such that $\mathbf{d}(\mathbb{k})$ is always defined. The coherent plus background noise signal vector $\widetilde{\mathbf{v}}$ is defined as $\widetilde{\mathbf{v}}(\mathbb{k}) = \widetilde{\mathbf{v}}_{\mathrm{c}}(\mathbb{k}) + \widetilde{\mathbf{v}}_{\mathrm{nc}}(\mathbb{k})$.

The eigenbeams $\widetilde{X}_{lm}$ are coherent across l and m [55,56], therefore the desired signal vector $\widetilde{\mathbf{x}}(\mathbb{k})$ can be expressed as $\widetilde{\mathbf{x}}(\mathbb{k}) = \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(\mathbb{k})\widetilde{X}_{00}(\mathbb{k})$, where

$$\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(\mathbb{k}) = \frac{\mathrm{E}\left[\widetilde{\mathbf{x}}(\mathbb{k})\widetilde{X}^{*}_{00}(\mathbb{k})\right]}{\mathrm{E}\left[|\widetilde{X}_{00}(\mathbb{k})|^{2}\right]} \tag{5.6}$$

is the partially normalized [with respect to $\widetilde{X}_{00}(\mathbb{k})$] coherence vector between $\widetilde{\mathbf{x}}(\mathbb{k})$ and $\widetilde{X}_{00}(\mathbb{k})$, and $\mathrm{E}[\cdot]$ denotes mathematical expectation. Using (5.6), the signal model in (5.5) can be rewritten as

$$\widetilde{\mathbf{p}}(\mathbb{k}) = \boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(\mathbb{k})\widetilde{X}_{00}(\mathbb{k}) + \widetilde{\mathbf{v}}(\mathbb{k}). \tag{5.7}$$

As $X(\mathbb{k}, \mathbf{r}_q)$, $V_{\mathrm{c}}(\mathbb{k}, \mathbf{r}_q)$ and $V_{\mathrm{nc}}(\mathbb{k}, \mathbf{r}_q)$ are mutually uncorrelated, and the spherical Fourier transform and division by the mode strength are linear operations, $\widetilde{X}_{lm}(\mathbb{k})$, $\widetilde{V}_{lm,\mathrm{c}}(\mathbb{k})$ and $\widetilde{V}_{lm,\mathrm{nc}}(\mathbb{k})$ are also mutually uncorrelated. The PSD matrix $\boldsymbol{\Phi}_{\widetilde{\mathbf{p}}}$ of $\widetilde{\mathbf{p}}$ can therefore be expressed as

$$\begin{aligned}
\boldsymbol{\Phi}_{\widetilde{\mathbf{p}}}(\mathbb{k}) &= \mathrm{E}\left[\widetilde{\mathbf{p}}(\mathbb{k})\widetilde{\mathbf{p}}^{\mathrm{H}}(\mathbb{k})\right] \\
&= \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}}(\mathbb{k}) + \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}(\mathbb{k}) \\
&= \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}}(\mathbb{k}) + \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}_{\mathrm{c}}}(\mathbb{k}) + \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}_{\mathrm{nc}}}(\mathbb{k}),
\end{aligned} \tag{5.8}$$

where

$$\begin{aligned}
\boldsymbol{\Phi}_{\widetilde{\mathbf{x}}}(\mathbb{k}) &= \mathrm{E}\left[\widetilde{\mathbf{x}}(\mathbb{k})\widetilde{\mathbf{x}}^{\mathrm{H}}(\mathbb{k})\right] = \phi_{\widetilde{X}_{00}}(\mathbb{k})\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(\mathbb{k})\boldsymbol{\gamma}^{\mathrm{H}}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(\mathbb{k}),\\
\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}(\mathbb{k}) &= \mathrm{E}\left[\widetilde{\mathbf{v}}(\mathbb{k})\widetilde{\mathbf{v}}^{\mathrm{H}}(\mathbb{k})\right] = \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}_{\mathrm{c}}}(\mathbb{k}) + \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}_{\mathrm{nc}}}(\mathbb{k}),\\
\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}_{\mathrm{c}}}(\mathbb{k}) &= \mathrm{E}\left[\widetilde{\mathbf{v}}_{\mathrm{c}}(\mathbb{k})\widetilde{\mathbf{v}}_{\mathrm{c}}^{\mathrm{H}}(\mathbb{k})\right] \text{ and}\\
\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}_{\mathrm{nc}}}(\mathbb{k}) &= \mathrm{E}\left[\widetilde{\mathbf{v}}_{\mathrm{nc}}(\mathbb{k})\widetilde{\mathbf{v}}_{\mathrm{nc}}^{\mathrm{H}}(\mathbb{k})\right]
\end{aligned}$$

are respectively the PSD matrices of $\widetilde{\mathbf{x}}(\mathbb{k}), \widetilde{\mathbf{v}}(\mathbb{k}), \widetilde{\mathbf{v}}_{\mathrm{c}}(\mathbb{k})$ and $\widetilde{\mathbf{v}}_{\mathrm{nc}}(\mathbb{k})$, $\phi_{\widetilde{X}_{00}}(\mathbb{k}) = \mathrm{E}\left[|\widetilde{X}_{00}(\mathbb{k})|^2\right]$ is the variance of $\widetilde{X}_{00}(\mathbb{k})$, and $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose.

Equation (5.7) contains the desired signal $\widetilde{X}_{00}(\mathbb{k})$ and is the basis for the design of our beamformer. The output $Z(\mathbb{k})$ of our beamformer is obtained by applying a complex weight $\mathbf{fi}^{\mathrm{H}}$ to each eigenbeam, and summing over all eigenbeams:

$$\begin{aligned}
Z(\mathbb{k}) &= \mathbf{fi}^{\mathrm{H}}(\mathbb{k})\widetilde{\mathbf{p}}(\mathbb{k})\\
&= \mathbf{fi}^{\mathrm{H}}(\mathbb{k})\widetilde{\mathbf{x}}(\mathbb{k}) + \mathbf{fi}^{\mathrm{H}}(\mathbb{k})\widetilde{\mathbf{v}}_{\mathrm{c}}(\mathbb{k}) + \mathbf{fi}^{\mathrm{H}}(\mathbb{k})\widetilde{\mathbf{v}}_{\mathrm{nc}}(\mathbb{k})\\
&= \widetilde{X}_{\mathrm{fd}}(\mathbb{k}) + \widetilde{V}_{\mathrm{rc}}(\mathbb{k}) + \widetilde{V}_{\mathrm{rnc}}(\mathbb{k}),
\end{aligned} \tag{5.9}$$

where $\widetilde{X}_{\mathrm{fd}}(\mathbb{k}) = \mathbf{fi}^{\mathrm{H}}(\mathbb{k})\widetilde{\mathbf{x}}(\mathbb{k}) = \mathbf{fi}^{\mathrm{H}}(\mathbb{k})\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(\mathbb{k})\widetilde{X}_{00}(\mathbb{k})$ is the filtered desired signal, $\widetilde{V}_{\mathrm{rc}}(\mathbb{k}) = \mathbf{fi}^{\mathrm{H}}(\mathbb{k})\widetilde{\mathbf{v}}_{\mathrm{c}}(\mathbb{k})$ is the residual coherent noise and $\widetilde{V}_{\mathrm{rnc}}(\mathbb{k}) = \mathbf{fi}^{\mathrm{H}}(\mathbb{k})\widetilde{\mathbf{v}}_{\mathrm{nc}}(\mathbb{k})$ is the residual background noise.

We now define two performance measures that will be used to derive our tradeoff beamformer. The first is the **noise reduction factor**, which measures the amount of noise attenuated by the beamformer [12], and is given by the ratio of the power of the noise at $\mathcal{M}_{\mathrm{ref}}$ to the power of the residual noise at the beamformer output. We define the narrowband noise reduction factor as

$$\xi_{\mathrm{nr}}\left[\mathbf{fi}(\mathbb{k})\right] = \frac{\phi_{\widetilde{V}_{00}}(\mathbb{k})}{\phi_{\widetilde{V}_{\mathrm{r}}}(\mathbb{k})} = \frac{\phi_{\widetilde{V}_{00}}(\mathbb{k})}{\mathbf{fi}^{\mathrm{H}}(\mathbb{k})\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}(\mathbb{k})\mathbf{fi}(\mathbb{k})}, \tag{5.10}$$

where $\phi_{\widetilde{V}_{00}}(\mathbb{k}) = \mathrm{E}\left[|\widetilde{V}_{00,\mathrm{c}}(\mathbb{k})|^2\right] + \mathrm{E}\left[|\widetilde{V}_{00,\mathrm{nc}}(\mathbb{k})|^2\right]$ is the variance of $\widetilde{V}_{00}(\mathbb{k})$ and $\phi_{\widetilde{V}_r}(\mathbb{k}) = \mathrm{E}\left[|\widetilde{V}_{\mathrm{rc}}(\mathbb{k})|^2\right] + \mathrm{E}\left[|\widetilde{V}_{\mathrm{rnc}}(\mathbb{k})|^2\right]$ is the variance of the residual noise.

The second measure is the **speech distortion index**, which measures the distortion of the desired speech signal $\widetilde{X}_{00}(\mathbb{k})$ introduced by the beamformer. The narrowband speech distortion index [12] is defined as

$$v_{\mathrm{sd}}\left[\mathbf{fi}(\mathbb{k})\right] = \frac{\mathrm{E}\left[|\widetilde{X}_{\mathrm{fd}}(\mathbb{k}) - \widetilde{X}_{00}(\mathbb{k})|^2\right]}{\phi_{\widetilde{X}_{00}}(\mathbb{k})} \tag{5.11a}$$

$$= \left|\mathbf{fi}^{\mathrm{H}}(\mathbb{k})\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(\mathbb{k}) - 1\right|^2. \tag{5.11b}$$

A tradeoff beamformer that achieves noise reduction while minimizing the speech distortion can then be designed according to the following optimization criteria [10,107]:

$$\min_{\mathbf{fi}(\mathbb{k})} v_{\mathrm{sd}}\left[\mathbf{fi}(\mathbb{k})\right] \text{ s.t. } \xi_{\mathrm{nr}}\left[\mathbf{fi}(\mathbb{k})\right] = \beta^{-1}$$

$$\min_{\mathbf{fi}(\mathbb{k})} \left|\mathbf{fi}^{\mathrm{H}}(\mathbb{k})\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(\mathbb{k}) - 1\right|^2 \text{ s.t. } \mathbf{fi}^{\mathrm{H}}(\mathbb{k})\Phi_{\widetilde{\mathbf{v}}}(\mathbb{k})\mathbf{fi}(\mathbb{k}) = \beta\phi_{\widetilde{V}_{00}}(\mathbb{k}),$$

where $0 < \beta < 1$ controls the level of noise reduction. Using a Lagrange multiplier, $\mu(\mathbb{k}) \geq 0$, to adjoin the constraint to the cost function, we deduce the tradeoff filter [10]:

$$\begin{aligned}\mathbf{fi}_{\mathrm{T},\mu}(\mathbb{k}) &= \phi_{\widetilde{X}_{00}}(\mathbb{k})\left[\boldsymbol{\Phi}_{\widetilde{\mathbf{x}}}(\mathbb{k}) + \mu(\mathbb{k})\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}(\mathbb{k})\right]^{-1}\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(\mathbb{k}) \\ &= \frac{\phi_{\widetilde{X}_{00}}(\mathbb{k})\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(\mathbb{k})\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(\mathbb{k})}{\mu(\mathbb{k}) + \phi_{\widetilde{X}_{00}}(\mathbb{k})\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}^{\mathrm{H}}(\mathbb{k})\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}}^{-1}(\mathbb{k})\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(\mathbb{k})},\end{aligned} \tag{5.12}$$

where the Lagrange multiplier, $\mu(\mathbb{k})$, satisfies the constraint $\xi_{\mathrm{nr}}\left[\mathbf{fi}(\mathbb{k})\right] = \beta^{-1}$. In the spatial domain, the tradeoff filter in (5.12) is also known as a *speech distortion weighted multichannel Wiener filter* (SDW-MWF) [31,107].

In practice, it is not easy to determine the optimal $\mu(\mathbb{k})$ and remove the dependency of the filter weights on $\mu(\mathbb{k})$, therefore $\mu(\mathbb{k})$ is chosen in an ad-hoc way and referred to as a *tradeoff parameter*. Increasing the value of $\mu(\mathbb{k})$ increases noise reduction at the expense of higher speech distortion. It has been shown [10] that for $\mu = 0$, this

corresponds to a SHD MVDR beamformer, while for $\mu = 1$, this corresponds to a SHD Wiener filter. The tradeoff parameter can be signal-dependent; for example, in [83] the authors used the SPP to increase the noise reduction when speech is likely to be absent.

5.3 Signal statistics estimation

In order to compute the tradeoff filter in (5.12), we must estimate the noise PSD matrix $\mathbf{\Phi}_{\widetilde{\mathbf{v}}}$, as well as the coherence vector $\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}$. Many techniques exist to estimate these statistics using the SPP in the spatial domain [23, 24, 105, 111]; in this section, we explain how to estimate them in the SHD using the DSPP. It should be noted that while in this chapter a tradeoff filter is used to extract the desired signal, the algorithm presented in this section could also be applied to other filters whose weights depend on the noise PSD matrix $\mathbf{\Phi}_{\widetilde{\mathbf{v}}}$ and/or the coherence vector $\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}$.

Due to the sparsity of speech in the time-frequency domain, it is commonly assumed that in a sound field comprising a mixture of speech sources, only one of them is active in each time-frequency bin [14, 87], i.e., that the sources are perfectly W-disjoint orthogonal. It has been shown that this is a reasonable approximation if the STFT window parameters are chosen appropriately [96].

For the purposes of the statistics estimation we therefore assume that, in a single time-frequency bin, only a single coherent source is active, whether it be the desired source or an interfering source. Although in practice this assumption does not always hold, particularly when multiple interfering speakers are present (see Sec. 5.6.2), only the desired source or the interfering sources are usually dominant in any one time-frequency bin, such that the resulting errors in the estimated statistics only have a small effect on the beamformer output (see Sec. 5.6.3). It should be noted that the tradeoff filter makes no such assumption, and can handle any number of simultaneously active sources.

Based on this assumption, we can then consider the following hypotheses regarding

the presence of desired speech and interference in each time-frequency bin:

$$\mathcal{H}_0(\mathfrak{k},\ell) : \widetilde{\mathbf{p}}(\mathfrak{k},\ell) = \widetilde{\mathbf{v}}_{\mathrm{nc}}(\mathfrak{k},\ell) \text{ indicating } \textit{speech absence};$$
$$\mathcal{H}_{1,\mathrm{c}}(\mathfrak{k},\ell) : \widetilde{\mathbf{p}}(\mathfrak{k},\ell) = \widetilde{\mathbf{v}}_{\mathrm{c}}(\mathfrak{k},\ell) + \widetilde{\mathbf{v}}_{\mathrm{nc}}(\mathfrak{k},\ell) \text{ indicating } \textit{interfering speech presence};$$
$$\mathcal{H}_{1,\mathrm{d}}(\mathfrak{k},\ell) : \widetilde{\mathbf{p}}(\mathfrak{k},\ell) = \widetilde{\mathbf{x}}(\mathfrak{k},\ell) + \widetilde{\mathbf{v}}_{\mathrm{nc}}(\mathfrak{k},\ell) \text{ indicating } \textit{desired speech presence}.$$

We define $\mathcal{H}_1 = \mathcal{H}_{1,c} \cup \mathcal{H}_{1,d}$, i.e. $\mathcal{H}_1$ indicates *speech presence* (desired or interfering). The signal $\widetilde{\mathbf{x}}$ originates from a source located *inside* the region of interest $\mathcal{R}$, while the signal $\widetilde{\mathbf{v}}_{\mathrm{c}}$ originates from a single source located *outside* $\mathcal{R}$.

5.3.1 Noise PSD matrix estimation

A minimum mean square error estimate of the noise PSD matrix taking into account the probability of these hypotheses is given by[3]

$$\begin{aligned}\mathrm{E}\left[\widetilde{\mathbf{v}}\widetilde{\mathbf{v}}^{\mathrm{H}}|\widetilde{\mathbf{p}}\right] =& \Pr\left[\mathcal{H}_0 \cup \mathcal{H}_{1,\mathrm{c}}|\widetilde{\mathbf{p}}\right] \mathrm{E}\left[\widetilde{\mathbf{v}}\widetilde{\mathbf{v}}^{\mathrm{H}}|\widetilde{\mathbf{p}}, \mathcal{H}_0 \cup \mathcal{H}_{1,\mathrm{c}}\right] \\ &+ \Pr\left[\mathcal{H}_{1,\mathrm{d}}|\widetilde{\mathbf{p}}\right] \mathrm{E}\left[\widetilde{\mathbf{v}}\widetilde{\mathbf{v}}^{\mathrm{H}}|\widetilde{\mathbf{p}}, \mathcal{H}_{1,\mathrm{d}}\right], \end{aligned} \tag{5.13}$$

where $\Pr\left[\mathcal{H}_{1,\mathrm{d}}|\widetilde{\mathbf{p}}\right]$ is the DSPP, $\Pr\left[\mathcal{H}_0 \cup \mathcal{H}_{1,\mathrm{c}}|\widetilde{\mathbf{p}}\right] = 1 - \Pr\left[\mathcal{H}_{1,\mathrm{d}}|\widetilde{\mathbf{p}}\right]$ is the desired speech absence probability, and $\mathrm{E}\left[\cdot|\cdot\right]$ denotes conditional expectation. A common way of approximating (5.13) is to recursively estimate the PSD matrix with a smoothing factor which depends on the SPP, as in [104, 111], such that the estimate is updated most rapidly when speech is absent.

The smoothing factor must be carefully chosen: if the noise PSD estimate is updated too rapidly, there is a risk that desired speech will leak into the estimate when the SPP is high, but not equal to 1, resulting in desired speech cancellation, whereas if the estimate is updated too slowly, non-stationary noise will not be effectively suppressed.

[3] For brevity, the dependencies on the discrete frequency and time indices $\mathfrak{k}$ and ℓ are omitted where possible in the following sections.

We would like to suppress a coherent speech source, which is non-stationary and has a similar spectral distribution to the desired speech (i.e., high energy at low frequencies). For this reason, we propose to estimate the PSD as

$$\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}(\ell) = \alpha'_{\mathrm{v}}\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}(\ell-1) + (1-\alpha'_{\mathrm{v}})\widetilde{\mathbf{p}}\,\widetilde{\mathbf{p}}^{\mathrm{H}}, \tag{5.14}$$

where

$$\alpha'_{\mathrm{v}} = \begin{cases} \alpha_{\mathrm{v}}, & \text{if } \Pr\left[\mathcal{H}_{1,\mathrm{d}}|\widetilde{\mathbf{p}}\right] < \Pr_{\mathrm{th}}; \\ 1, & \text{otherwise,} \end{cases} \tag{5.15}$$

and $0 < \alpha_{\mathrm{v}} \leq 1$ is a smoothing factor. The PSD estimate is therefore only updated if the DSPP is below a threshold $\Pr_{\mathrm{th}}$.

5.3.2 Coherence vector estimation

The coherence vector $\boldsymbol{\gamma}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}$ is given by the first column of $\mathbf{\Phi}_{\widetilde{\mathbf{x}}}$ divided by the first element $\phi_{\widetilde{X}_{00}}$, and is estimated by

$$\hat{\boldsymbol{\gamma}}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}} = \hat{\phi}^{-1}_{\widetilde{X}_{00}}\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}}\,\mathbf{i}_N, \tag{5.16}$$

where $\mathbf{i}_N = [1\ 0\ \cdots\ 0]^{\mathrm{T}}$ is a vector of length N. Since the noise is always present, the desired signal is not directly observable. Therefore, we propose to first compute an estimate of the desired speech plus background noise PSD $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}+\widetilde{\mathbf{v}}_{\mathrm{nc}}}$ as

$$\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}+\widetilde{\mathbf{v}}_{\mathrm{nc}}}(\ell) = \alpha'_{\mathrm{xv}_{\mathrm{nc}}}\widetilde{\mathbf{p}}\,\widetilde{\mathbf{p}}^{\mathrm{H}} + \left[1-\alpha'_{\mathrm{xv}_{\mathrm{nc}}}\right]\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}+\widetilde{\mathbf{v}}_{\mathrm{nc}}}(\ell-1), \tag{5.17}$$

where $\alpha'_{\mathrm{xv}_{\mathrm{nc}}} = \Pr\left[\mathcal{H}_{1,\mathrm{d}}|\widetilde{\mathbf{p}}\right](1-\alpha_{\mathrm{xv}_{\mathrm{nc}}})$ and $0 < \alpha_{\mathrm{xv}_{\mathrm{nc}}} \leq 1$ is a smoothing factor. We can now obtain an estimate $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}}$ of the desired speech PSD matrix using

$$\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}} = \hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}+\widetilde{\mathbf{v}}_{\mathrm{nc}}} - \hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}_{\mathrm{nc}}}. \tag{5.18}$$

The coherence vector estimate $\hat{\boldsymbol{\gamma}}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}$ is therefore updated most rapidly when desired speech is present.

We assume that the background noise $\widetilde{\mathbf{v}}_{\mathrm{nc}}$ is stationary; an estimate $\hat{\boldsymbol{\Phi}}_{\widetilde{\mathbf{v}}_{\mathrm{nc}}}$ of its PSD matrix can therefore be obtained during initial noise only frames. If the background noise is not stationary, its PSD matrix can be estimated using $\Pr\left[\mathcal{H}_0|\widetilde{\mathbf{p}}\right]$ in a similar way to the noise PSD matrix in (5.14).

5.4 Desired speech presence probability estimation

Using the definition of conditional probability, the DSPP $\Pr\left[\mathcal{H}_{1,\mathrm{d}}|\widetilde{\mathbf{p}}\right]$ can be expressed as

$$\begin{aligned}\Pr\left[\mathcal{H}_{1,\mathrm{d}}|\widetilde{\mathbf{p}}\right] &= \Pr\left[\mathcal{H}_{1,\mathrm{d}} \cap \mathcal{H}_1|\widetilde{\mathbf{p}}\right] \\ &= \Pr\left[\mathcal{H}_{1,\mathrm{d}}|\mathcal{H}_1, \widetilde{\mathbf{p}}\right] \cdot \Pr\left[\mathcal{H}_1|\widetilde{\mathbf{p}}\right].\end{aligned}$$

The term $\Pr\left[\mathcal{H}_1|\widetilde{\mathbf{p}}\right]$ can be determined using a Gaussian model-based multichannel SPP estimator [104], while in this work we assume the term $\Pr\left[\mathcal{H}_{1,\mathrm{d}}|\mathcal{H}_1, \widetilde{\mathbf{p}}\right]$ can be approximated based on an instantaneous DOA estimate $\hat{\Omega}$, i.e.,

$$\Pr\left[\mathcal{H}_{1,\mathrm{d}}|\mathcal{H}_1, \widetilde{\mathbf{p}}\right] \approx \Pr[\mathcal{H}_{1,\mathrm{d}}|\mathcal{H}_1, \hat{\Omega}].$$

The multiplication of the SPP by a DOA-based probability allows us to differentiate between desired coherent sources and interfering coherent sources. The combination of these two probabilities, along with the method for estimating the DOA-based probability, are the two main contributions of this chapter. In the following, we explain how to estimate the SPP and DOA-based probability.

5.4.1 Multichannel speech presence probability

Assuming the desired speech, coherent noise, and background noise can be modeled as complex multivariate Gaussian random variables, an *a posteriori* multichannel SPP estimate is given by [104]:

$$\Pr\left[\mathcal{H}_1|\widetilde{\mathbf{p}}\right] = \left\{1 + \frac{1-\rho}{\rho}(1+\xi)e^{-\frac{\beta}{1+\xi}}\right\}^{-1}, \tag{5.19}$$

where $\rho = \Pr\left[\mathcal{H}_1\right]$ denotes the *a priori* SPP, β is defined as

$$\beta = \widetilde{\mathbf{p}}^{\mathrm{H}}\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}_{\mathrm{nc}}}^{-1}\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{r}}}\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}_{\mathrm{nc}}}^{-1}\widetilde{\mathbf{p}}, \tag{5.20}$$

and ξ is defined as

$$\xi = \mathrm{tr}\left(\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}_{\mathrm{nc}}}^{-1}\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{r}}}\right). \tag{5.21}$$

The PSD matrix $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{r}}}$ is given by

$$\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{r}}} = \hat{\mathbf{\Phi}}_{\widetilde{\mathbf{p}}} - \hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}_{\mathrm{nc}}}, \tag{5.22}$$

and represents the desired signal plus coherent noise. ThePSD matrix $\mathbf{\Phi}_{\widetilde{\mathbf{p}}}$ is recursively estimated as

$$\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{p}}}(\ell) = \alpha_{\mathrm{p}}\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{p}}}(\ell-1) + (1-\alpha_{\mathrm{p}})\widetilde{\mathbf{p}}\,\widetilde{\mathbf{p}}^{\mathrm{H}}, \tag{5.23}$$

where $0 < \alpha_{\mathrm{p}} \leq 1$ is a smoothing factor.

5.4.2 DOA-based probability

The DOA-based probability $\Pr[\mathcal{H}_{1,\mathrm{d}}|\mathcal{H}_1, \hat{\Omega}]$ is obtained from the instantaneous DOA estimates and the associated uncertainty. Under specific conditions (e.g., direct-to-reverberant ratio, signal-to-noise ratio), we can find an empirical probability distribution function (PDF) $f(\hat{\Omega}|\Omega;\Sigma)$ that describes the distribution of the DOA estimates $\hat{\Omega}$ ob-

tained using a specific narrowband DOA estimation algorithm for a source at a DOA $\Omega = (\theta, \phi)$.

A training phase is used to estimate this empirical PDF. An analytic PDF is then fitted to the estimated DOAs for each specific condition. The PDF is denoted by $f(\hat{\Omega}|\Omega, \Sigma)$ where Σ describes the uncertainty associated with the estimate of Ω. A region of interest is defined by a function $R(\Omega)$, where $0 \leq R(\Omega) \leq 1$. The DOA-based probability is then given by

$$\begin{aligned} \Pr[\mathcal{H}_{1,\mathrm{d}}|\mathcal{H}_1, \hat{\Omega}] &= \Pr[\Omega \in \mathcal{R}|\hat{\Omega}] && \text{(5.24a)} \\ &= \int_{\Omega \in \mathcal{R}} f(\Omega|\hat{\Omega}; \Sigma) \mathrm{d}\Omega && \text{(5.24b)} \\ &= \int_{\Omega \in \mathcal{R}} \frac{f(\hat{\Omega}|\Omega; \Sigma) f(\Omega)}{f(\hat{\Omega})} \mathrm{d}\Omega, && \text{(5.24c)} \end{aligned}$$

where $\mathrm{d}\Omega = \sin\theta \mathrm{d}\theta \mathrm{d}\phi$ and we have used Bayes' rule to go from (5.24b) to (5.24c). The marginal PDF $f(\Omega)$ can be modeled using *a priori* information on possible source positions, while the marginal PDF $f(\hat{\Omega})$ can be estimated by observing the DOA estimates during the training phase.

In this work, the DOA is estimated using the pseudointensity vector method [57] (as presented in Sec. 4.1). The pseudointensity vector $\mathbf{I}$ is conceptually similar to the acoustic intensity vector, which describes the magnitude and direction of the transport of acoustic energy, but instead of being computed using particle velocity measurements [27, 28], it is computed using the zero- and first-order eigenbeams P_{00}, $P_{1(-1)}$, P_{10} and P_{11} obtained with a spherical microphone array [57][4]:

$$\mathbf{I} = \frac{1}{2} \Re \left\{ \widetilde{P}_{00}^{*} \begin{bmatrix} \sum_{m=-1}^{1} \widetilde{P}_{1m} \, Y_{1m}(\frac{\pi}{2}, \pi) \\ \sum_{m=-1}^{1} \widetilde{P}_{1m} \, Y_{1m}(\frac{\pi}{2}, -\frac{\pi}{2}) \\ \sum_{m=-1}^{1} \widetilde{P}_{1m} \, Y_{1m}(\pi, 0) \end{bmatrix} \right\}, \tag{5.25}$$

[4] It should be noted that although the dependency on the discrete frequency index $\breve{k}$ has been omitted, the DOA-based probability and pseudointensity vector are frequency-dependent.

where $\Re\{\cdot\}$ denotes the real part of a complex number. An estimate $\hat{\mathbf{u}}$ of the unit vector $\mathbf{u}$ with direction Ω is given by

$$\hat{\mathbf{u}}(\ell) = -\frac{\sum_{\ell'=\ell-\tau+1}^{\ell} \mathbf{I}(\ell')}{\| \sum_{\ell'=\ell-\tau+1}^{\ell} \mathbf{I}(\ell') \|_2}. \tag{5.26}$$

By summing the pseudointensity vectors over τ time frames, we give a higher weight to pseudointensity vectors with a high norm, which are considered to be more reliable. Finally, the instantaneous DOA estimate $\hat{\Omega}$ is given by the direction of the vector $\hat{\mathbf{u}}$. The accuracy of the DOA estimates obtained using this method is evaluated in Sec. 4.1.5.

The DOA estimates obtained using the pseudointensity vector method can be represented by the Fisher distribution [39], a probability distribution on the sphere with two parameters: the mean direction and the concentration parameter κ. The Fisher distribution is rotationally symmetric about the mean direction, which is assumed to be the true DOA Ω. The concentration parameter κ can be considered to be independent of Ω due to spherical symmetry, providing the source and array are reasonably far from the room boundaries, and is estimated during the training phase using the method described in [108].

Using the Fisher distribution, the PDF $f(\hat{\Omega}|\Omega, \Sigma)$ is then given by [39, 74]

$$f(\hat{\Omega}|\Omega, \kappa) = \frac{\kappa}{4\pi \sinh \kappa} e^{\kappa\, \mathbf{u}^{\mathrm{T}}\hat{\mathbf{u}}} \tag{5.27a}$$

$$= \frac{\kappa}{2\pi \left(e^{\kappa} - e^{-\kappa}\right)} e^{\kappa\, \mathbf{u}^{\mathrm{T}}\hat{\mathbf{u}}}. \tag{5.27b}$$

Due to the symmetry of the distribution about Ω, the PDF only depends on κ and $\mathbf{u}^{\mathrm{T}}\hat{\mathbf{u}}$, i.e., the cosine of the angle between the true and estimated DOAs (Ω and $\hat{\Omega}$, respectively), which we will call the *opening angle*. As κ increases, the distribution of $\hat{\Omega}$ becomes more concentrated around Ω, or equivalently the distribution of the opening angles becomes more concentrated around 0. We also note that for this choice of distribution, $f(\hat{\Omega}|\Omega, \kappa) = f(\Omega|\hat{\Omega}, \kappa)$, and as a result the DOA-based probability can be computed

from (5.24b) without estimating $f(\hat{\Omega})$ and $f(\Omega)$.

5.5 Algorithm summary

The noise PSD matrix $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}$ and coherence vector $\hat{\boldsymbol{\gamma}}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}$ are recursively estimated for all frequency indices k according to the following steps:

1. Estimate $\Pr[\mathcal{H}_{1,\mathrm{d}}(\ell)|\mathcal{H}_1(\ell), \hat{\Omega}(\ell)]$, the DOA-based probability:
   (a) Compute the pseudointensity vector $\mathbf{I}(\ell)$ using (5.25).
   (b) Compute the unit vector $\hat{\mathbf{u}}(\ell)$ using $\mathbf{I}(\ell), \mathbf{I}(\ell-1), \ldots, \mathbf{I}(\ell-\tau+1)$ and (5.26).
   (c) Compute the PDF $f(\hat{\Omega}|\Omega, \kappa)$ using $\hat{\mathbf{u}}(\ell)$, the concentration parameter κ estimated during the training phase and (5.27b).
   (d) Estimate $\Pr[\mathcal{H}_{1,\mathrm{d}}(\ell)|\mathcal{H}_1(\ell), \hat{\Omega}(\ell)]$ using $f(\hat{\Omega}|\Omega, \kappa)$ and (5.24b).
2. Update $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{p}}}(\ell)$ using (5.23).
3. Estimate $\mathbf{\Phi}_{\widetilde{\mathbf{r}}}(\ell)$ as $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{r}}}(\ell) = \hat{\mathbf{\Phi}}_{\widetilde{\mathbf{p}}}(\ell) - \hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}_{\mathrm{nc}}}$, where $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}_{\mathrm{nc}}}$ is estimated during initial background noise only frames.
4. Estimate the *a posteriori* **multichannel SPP** $\Pr[\mathcal{H}_1(\ell)|\widetilde{\mathbf{p}}(\ell)]$ according to (5.19), (5.20) and (5.21), using $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{r}}}(\ell)$ and $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}_{\mathrm{nc}}}$.
5. Compute the DSPP $\Pr[\mathcal{H}_{1,\mathrm{d}}(\ell)|\widetilde{\mathbf{p}}(\ell)]$ as the product of $\Pr[\mathcal{H}_{1,\mathrm{d}}(\ell)|\mathcal{H}_1(\ell), \hat{\Omega}(\ell)]$ and $\Pr[\mathcal{H}_1(\ell)|\widetilde{\mathbf{p}}(\ell)]$.
6. Update $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}(\ell)$ according to (5.14) by using $\Pr[\mathcal{H}_{1,\mathrm{d}}(\ell)|\widetilde{\mathbf{p}}(\ell)]$.
7. Update $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}+\widetilde{\mathbf{v}}_{\mathrm{nc}}}(\ell)$ according to (5.17) by using $\Pr[\mathcal{H}_{1,\mathrm{d}}(\ell)|\widetilde{\mathbf{p}}(\ell)]$, and compute $\hat{\boldsymbol{\gamma}}_{\widetilde{\mathbf{x}}\widetilde{X}_{00}}(\ell)$ according to (5.16).

In Fig. 5.1, the complete noise reduction algorithm is summarized in the form of a block diagram. The gray blocks refer to the steps in the algorithm summary above.

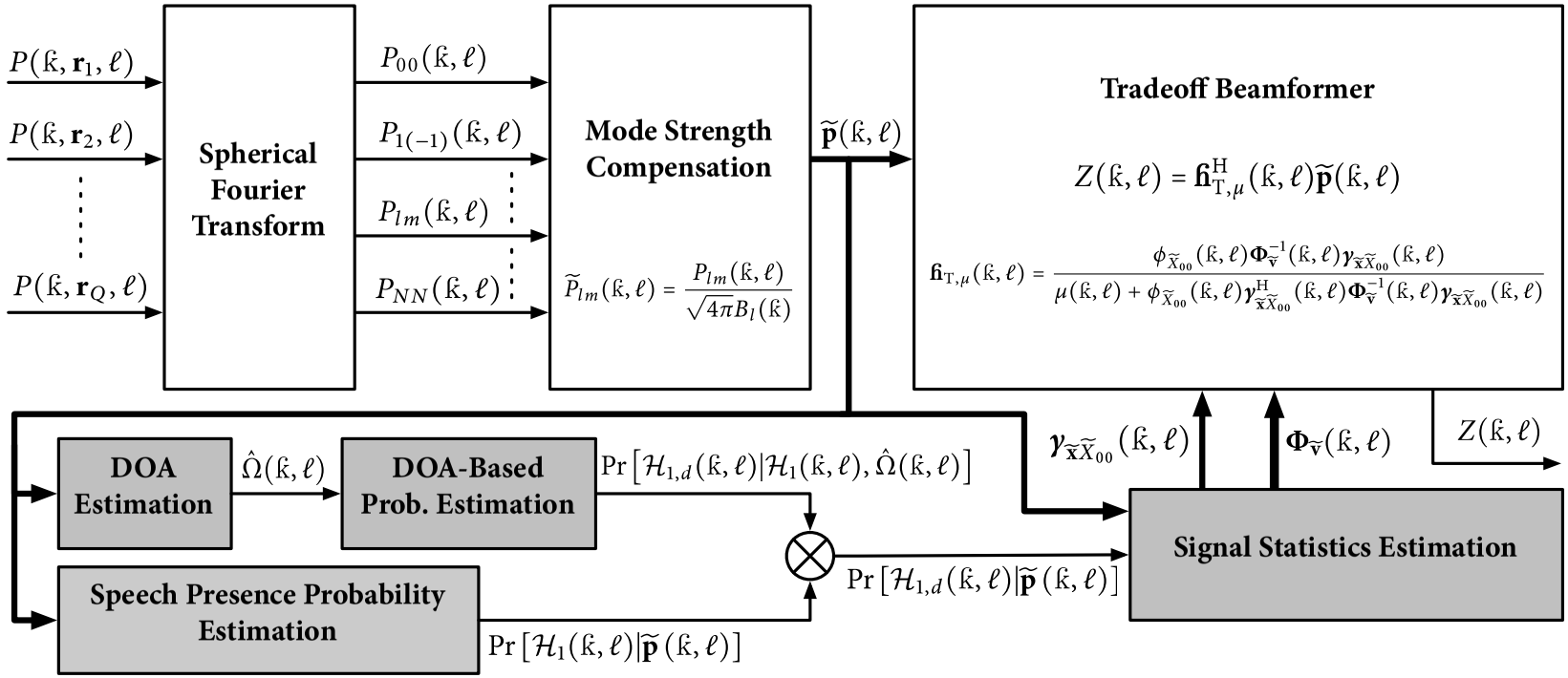


Figure 5.1: Block diagram of the complete noise reduction algorithm, including the tradeoff beamformer and DOA-based statistics estimation. The gray blocks refer to the steps in the algorithm summary (Sec. 5.5).

5.6 Performance evaluation

The evaluation of the performance of the proposed noise reduction algorithm consists of two parts: 1) evaluating the proposed DSPP estimation method described in Sec. 5.4, which is used in the estimation of the desired speech statistics, and 2) evaluating the performance of the tradeoff beamformer. An evaluation of the performance of the tradeoff beamformer under the assumption that the signal statistics are perfectly estimated can be found in an earlier contribution [56].

As mentioned in Sec. 5.2, in previous work the tradeoff parameter μ has been chosen to be a function of the SPP [83]. In this work, we make μ a function of the DSPP, i.e.,

$$\mu(k,\ell) = \frac{1}{\eta\frac{1}{\mu'} + (1-\eta)\Pr\left[\mathcal{H}_{1,\mathrm{d}}(k,\ell)|\widetilde{\mathbf{p}}(k,\ell)\right]}, \tag{5.28}$$

where $0 \leq \eta \leq 1$ and $\mu > 0$. As η decreases, the influence of the DSPP on the tradeoff parameter μ increases. For $\eta = 1$, μ is fixed and equals μ', whereas for $\eta = 0$, μ is equal to the inverse of the DSPP.

5.6.1 Experimental setup

In order to evaluate the proposed algorithm, clean speech signals were convolved with measured acoustic impulse responses (AIRs) from one of the laboratories at Fraunhofer IIS [103]. The reverberation time of the room with dimensions $7.5 \times 9.3 \times 4.2$ m was $T_{60} \approx 330$ ms. The AIRs were measured using an *Eigenmike* [79], i.e., a $Q = 32$ microphone rigid spherical array with radius 4.2 cm, located in the centre of the room. The desired talker was located at an inclination and azimuth of approximately $(95^\circ, 175^\circ)$, respectively, and a distance of 1.8 m from the centre of the array. The first interfering talker was located at approximately $(95^\circ, 115^\circ)$ and a distance of 2.3 m from the array centre, and the second interfering talker at $(40^\circ, 0^\circ)$ and a distance of 3.0 m from the array centre.

The desired and interfering speech signals consisted of male and female speech from the EBU SQAM dataset [37]. Four consecutive 15 s segments were used in the

evaluation: desired speaker only, single interfering speaker, desired speaker and single interfering speaker, and desired speaker and two interfering speakers. The background noise consisted of spatio-temporally white Gaussian noise with a constant input signal to incoherent noise ratio (iSINR) of 25 dB at $\mathcal{M}_{\text{ref}}$. It should be noted that the incoherent noise power at $\mathcal{M}_{\text{ref}}$ is reduced by a factor of $Q|B_0(\mathring{k})|^2$ with respect to its power at the microphones [55]; at low frequencies, where $B_0(\mathring{k})$ is lowest, the incoherent noise power is approximately 15 dB lower at $\mathcal{M}_{\text{ref}}$ than at the microphones[5]. The coherent noise power was set in order to obtain a given input signal to coherent noise ratio (iSCNR) at $\mathcal{M}_{\text{ref}}$, taking into account only frames where *both* interfering talkers were active according to ITU-T Rec. P.56 [53]. The local iSCNR was therefore higher in frames where only one interfering talker was active. Coherent and incoherent noise levels were set based on active speech levels, computed according to ITU-T Rec. P.56 [53].

The processing was performed in the STFT domain at a sampling frequency of 8 kHz with a frame length of 64 ms and a 50% overlap between successive frames, as in [105]. The beamformer was applied to eigenbeams of order up to $L = 3$, resulting in a total of $N = (L+1)^2 = 16$ eigenbeams. In order to reduce the computational complexity, the multichannel SPP was estimated based only on zero- and first-order eigenbeams, highlighting an advantage of working in the SHD. The smoothing factors in (5.23), (5.15) and (5.17) were empirically chosen as $\alpha_{\text{p}} = 0.8$, $\alpha_{\text{v}} = 0.7$ (with $\text{Pr}_{\text{th}} = 0.01$) and $\alpha_{\text{xv}_{\text{nc}}} = 0.9$, respectively, in order to achieve high noise reduction and low speech distortion. The *a priori* SPP ρ was fixed to 0.4, as in [104], and the pseudointensity vectors used in the DOA estimation were averaged over $\tau = 4$ time frames.

5.6.2 Desired speech presence probability

In the following, we evaluate the results of the DSPP estimation described in Sec. 5.4. In order to compute the DOA-based probability in (5.24b), the distribution $f(\Omega \,|\, \hat{\Omega}, \Sigma)$ is

[5] The iSINR at the sensors is therefore relatively low. The choice of iSINR is made to demonstrate that the proposed algorithm is robust to high levels of sensor noise. A higher iSINR could be chosen, and would show improved SPP estimation.

required. The distribution $f(\Omega \,|\, \hat{\Omega}, \Sigma)$ was modeled by a Fisher distribution where the uncertainty parameter Σ is given by a concentration parameter κ that is estimated during a training phase. The training was done using AIRs simulated with SMIRgen [54, 62], an AIR generator for spherical arrays based on the algorithm presented in Chapter 3. The reverberation time, source-array distance and iSINR chosen for the training were the same as in Sec. 5.6.1, such that the training conditions were similar to those where the tradeoff beamformer was applied. The integral in (5.24b) was evaluated numerically over a region $\mathcal{R}$ centred around the desired source's true DOA, defined as

$$\Omega = (\theta, \phi) \in \mathcal{R} \text{ if } \theta \in [80°, 110°] \text{ and } \phi \in [160°, 190°].$$

Particularly in the presence of strong early reflections, the DOA estimates at a given frequency might not be centred around the true DOA, i.e., can be biased. In order to reduce the bias and estimate a meaningful concentration parameter κ that will hold for all DOAs, we combine the DOA estimates obtained by varying the source-array positions (5 different positions) and keeping the rest of the training conditions (true DOA, source-array distance, reverberation time and iSINR) fixed. We then estimate the concentration parameter based on this multimodal distribution. Note that due to the frequency-dependence of the array's directivity, the concentration parameter must be estimated for each frequency. In Fig. 5.2 the estimated DOAs are plotted for two different frequencies. As expected, the DOA estimates have a lower concentration at low frequencies, where the array has lower directivity.

In Fig. 5.3 we plot some illustrative time-frequency plots of the opening angles between the true DOA of the desired source and the estimated DOAs [Fig. 5.3(a)], the DOA-based probability [Fig. 5.3(b)], the multichannel SPP [Fig. 5.3(c)], and the product of these two probabilities, the DSPP [Fig. 5.3(d)]. The results were obtained for an iSCNR of 0 dB at $\mathcal{M}_{\text{ref}}$. The signal was divided into four time segments, as in 5.6.1, namely, in the first segment, only the desired talker is active, then only a single interfering talker is

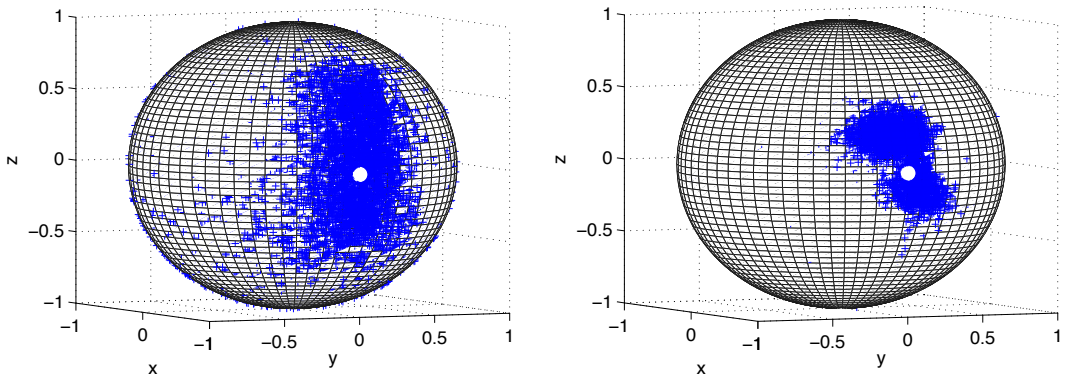


Figure 5.2: DOA estimates obtained using 5 source-array positions with identical true DOA (indicated by a white marker), at (a) 150 Hz and (b) 1.5 kHz.

active, then a desired talker and a single interfering talker are active, and finally a desired talker and two interfering talkers are active. It can be seen that multiplying the commonly used multichannel SPP by a DOA-based probability results in a sufficiently small DSPP when only interfering talkers are present. This allows us to distinguish between desired and undesired coherent sources, and derive accurate estimators for their respective PSD matrices, which are required to compute the tradeoff beamformer weights.

5.6.3 Tradeoff beamformer

In order to evaluate the performance of the combined DOA-based statistics estimation algorithm and tradeoff beamformer, we considered the following performance measures:

- ΔsegSNR, the **improvement in the segmental signal to noise ratio** (segSNR) with respect to the best sensor (i.e., the sensor with the best segSNR), where the signal-to-noise ratio (SNR) was given by the ratio of the power of the desired speech to the power of the coherent and background noise.

- segSDI, the **segmental speech distortion index**, as defined in [10, eqn 4.44] and [49, eqn 30], with respect to the desired speech signal at $\mathcal{M}_{\text{ref}}$. The segSDI is equal to 0 if there is no distortion, and is greater than 0 when distortion occurs.

- segBNRF, the **segmental background noise reduction factor**, given by the ratio

(a) Opening angles (degrees)

(b) $\Pr[\mathcal{H}_{1,\mathrm{d}}|\mathcal{H}_1, \hat{\Omega}]$

(c) $\Pr[\mathcal{H}_1|\widetilde{\mathbf{p}}]$

(d) $\Pr[\mathcal{H}_{1,\mathrm{d}}|\widetilde{\mathbf{p}}]$

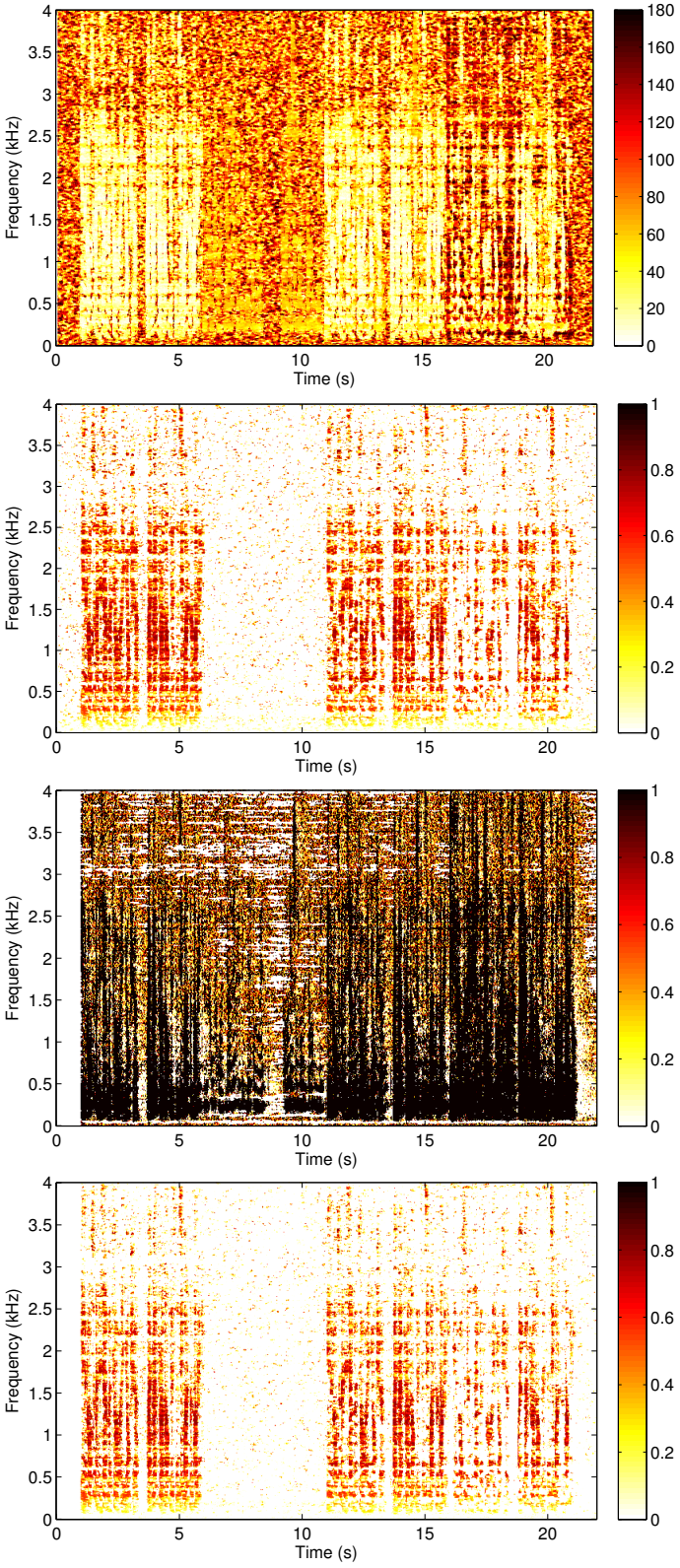


Figure 5.3: Time-frequency plots of (a) opening angles, (b) DOA-based probability $\Pr[\mathcal{H}_{1,\mathrm{d}}|\mathcal{H}_1, \hat{\Omega}]$, (c) *a posteriori* multichannel SPP $\Pr[\mathcal{H}_1|\widetilde{\mathbf{p}}]$, (d) DSPP $\Pr[\mathcal{H}_{1,\mathrm{d}}|\widetilde{\mathbf{p}}]$. The iSCNR was 0 dB at $\mathcal{M}_{\mathrm{ref}}$.

of the power of the background noise at the best sensor (i.e., the sensor with the lowest background noise power) to the power of the background noise at the output of the beamformer.

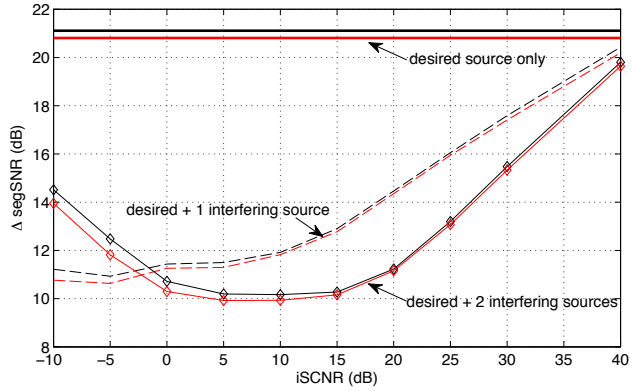
- segCNRF, the **segmental coherent noise reduction factor**, given by the ratio of the power of the coherent noise at $\mathcal{M}_{\text{ref}}$ to the power of the coherent noise at the output of the beamformer. It should be noted that the segCNRF with respect to the best sensor would be slightly lower, due to the fact that $\mathcal{M}_{\text{ref}}$ is omnidirectional, while the sensors have additional directivity provided by the rigid sphere.

All performance measures were computed in the time domain using non-overlapping frames of length 16 ms. The segSNR and segSDI were averaged over all frames that contained desired speech. The segCNRF was averaged over all frames that contained interfering speech. A frame was considered to contain speech if the average energy of the frame was at least −30 dB with respect to the frame with the highest average energy. The performance measures were averaged in the log domain, except for the speech distortion index, which was averaged in the linear domain.

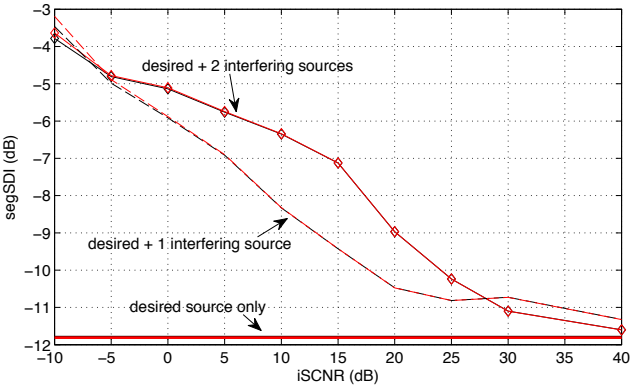
The performance measures are plotted in Fig. 5.4 as a function of the iSCNR. The performance was evaluated separately for each of the speech segments (desired source only, one interfering source only, desired source and one interfering source, desired source and two interfering sources). Two sets of tradeoff parameters were used: $\eta = 1$, $\mu' = 1$, resulting in a SHD Wiener filter and $\eta = 0.25$, $\mu' = 1$, resulting in a DSPP-based tradeoff parameter μ ranging from 1 to 4.

Although the tradeoff beamformer outperforms the SHD Wiener filter across all performance measures, in most cases the performance difference is quite small (0–1 dB). The largest difference is observed in the presence of a single interfering source, where the DSPP-based tradeoff parameter leads to stronger noise and interference reduction. Comparing the segCNRF and segBNRF curves in Fig. 5.4(c) and Fig. 5.4(d), a tradeoff between coherent and incoherent noise reduction can be observed, which is consistent

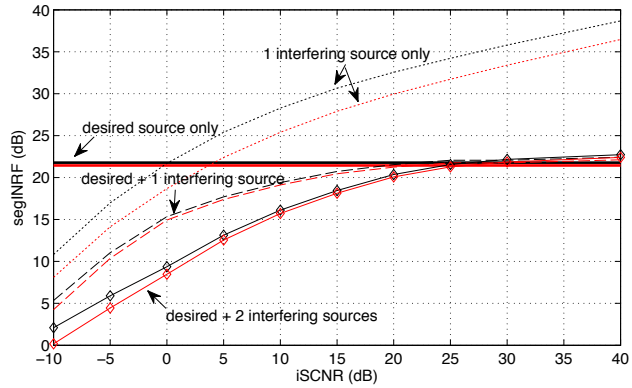
(a) Segmental SNR improvement ΔsegSNR (with respect to the best sensor)



(b) Segmental speech distortion index segSDI (with respect to $\mathcal{M}_{\text{ref}}$)



(c) Segmental background noise reduction factor segBNRF (with respect to the best sensor)



(d) Segmental coherent noise reduction factor segCNRF (with respect to $\mathcal{M}_{\text{ref}}$)

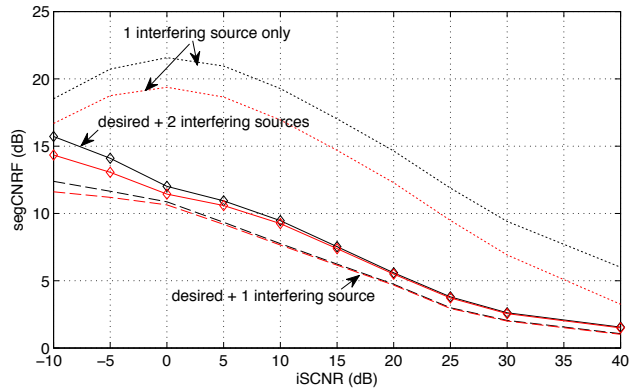


Figure 5.4: Performance measures as a function of the input signal to coherent noise ratio (iSCNR) at $\mathcal{M}_{\text{ref}}$, for two different parameter sets: $\eta = 1$, $\mu' = 1$ (red curves) and $\eta = 0.25$, $\mu' = 1$ (black curves).

with the findings presented in [48]. Finally, note that in all scenarios, as the iSCNR increases the performance of the beamformers converges to the performance when only a desired talker is present, as expected.

In Table 5.1, some performance measures are provided as a function of the parameters η and μ', which control the tradeoff parameter μ. In these results, it can again be observed that during interference-only periods, the DSPP-based tradeoff parameter ($\eta < 1$) leads to stronger noise reduction than the fixed tradeoff parameter ($\eta = 1$): when $\eta = 0$, i.e., μ is inversely proportional to the DSPP, the segCNRF and segBNRF are 13–18 dB higher than for $\eta = 1$. For the other scenarios, the extreme case of $\eta = 0$ once more shows the highest noise reduction, this time at the cost of increased speech distortion. However, in the rest of the cases, the speech distortion index is largely unaffected by changes in η and μ', indicating that speech distortion is mostly introduced by errors in the DSPP estimation and hence the signal statistics, rather than by an increase in the tradeoff parameter of the beamformer.

Sample spectrograms are presented in Fig. 5.5, for a fixed iSCNR of 0 dB at $\mathcal{M}_{\text{ref}}$. The sequence of speech segments is as described in 5.6.1, and each segment has duration 5 s. The spectra of the desired speech signal and the mixture are illustrated in Fig. 5.5(a) and Fig. 5.5(b), respectively. The spectrograms of the tradeoff beamformer output for two different tradeoff parameters are illustrated in Fig. 5.5(c) and Fig. 5.5(d). Choosing $\mu' = 0$ in Fig. 5.5(c) results in $\mu = 0$ which corresponds to the SHD MVDR beamformer. The effect of a DSPP-dependent tradeoff parameter is visible in Fig. 5.5(d), where the coherent noise reduction performance is improved compared to the SHD MVDR beamformer in Fig. 5.5(c). This effect is most visible in the interference-only segment, where a segCNRF improvement of about 5 dB is obtained.

Table 5.1: Tradeoff beamformer performance measures (in dB) as a function of the parameters η and μ', for three different scenarios.

a) Desired speaker only:

η	μ'	μ	segSDI	ΔsegSNR
0	> 0	DSPP^{-1}	−10.3	24.1
0.25	1	1 – 4	−11.8	21.1
	2	1.14 – 8	−11.8	21.4
0.5	1	1 – 2	−11.8	20.9
	2	1.33 – 4	−11.8	21.1
1	→ 0	0	−11.8	20.6
	1	1	−11.8	20.8
	2	2	−11.8	20.9

b) Desired speaker and single interfering speaker:

			iSCNR = 5 dB		iSCNR = 15 dB	
η	μ'	μ	segSDI	ΔsegSNR	segSDI	ΔsegSNR
0	> 0	DSPP^{-1}	−6.4	14.0	−8.6	14.8
0.25	1	1 – 4	−6.9	11.5	−9.4	12.9
	2	1.14 – 8	−6.9	11.6	−9.4	13.0
0.5	1	1 – 2	−6.9	11.4	−9.4	12.8
	2	1.33 – 4	−6.9	11.5	−9.4	12.9
1	→ 0	0	−6.9	11.2	−9.4	12.7
	1	1	−6.9	11.3	−9.4	12.8
	2	2	−6.9	11.4	−9.4	12.8

c) Single interfering speaker:

			iSCNR = 5 dB		iSCNR = 15 dB	
η	μ'	μ	segCNRF	segBNRF	segCNRF	segBNRF
0	> 0	DSPP^{-1}	30.7	38.4	27.8	42.6
0.25	1	1 – 4	25.4	22.7	17.0	30.7
	2	1.14 – 8	27.8	19.6	19.0	33.0
1	→ 0	0	17.3	20.8	13.5	26.5

(a) $\widetilde{X}_{00}$

(b) $\widetilde{P}_{00}$

(c) Z for $\eta = 1, \mu' = 0$

(d) Z for $\eta = 0.25, \mu' = 2$

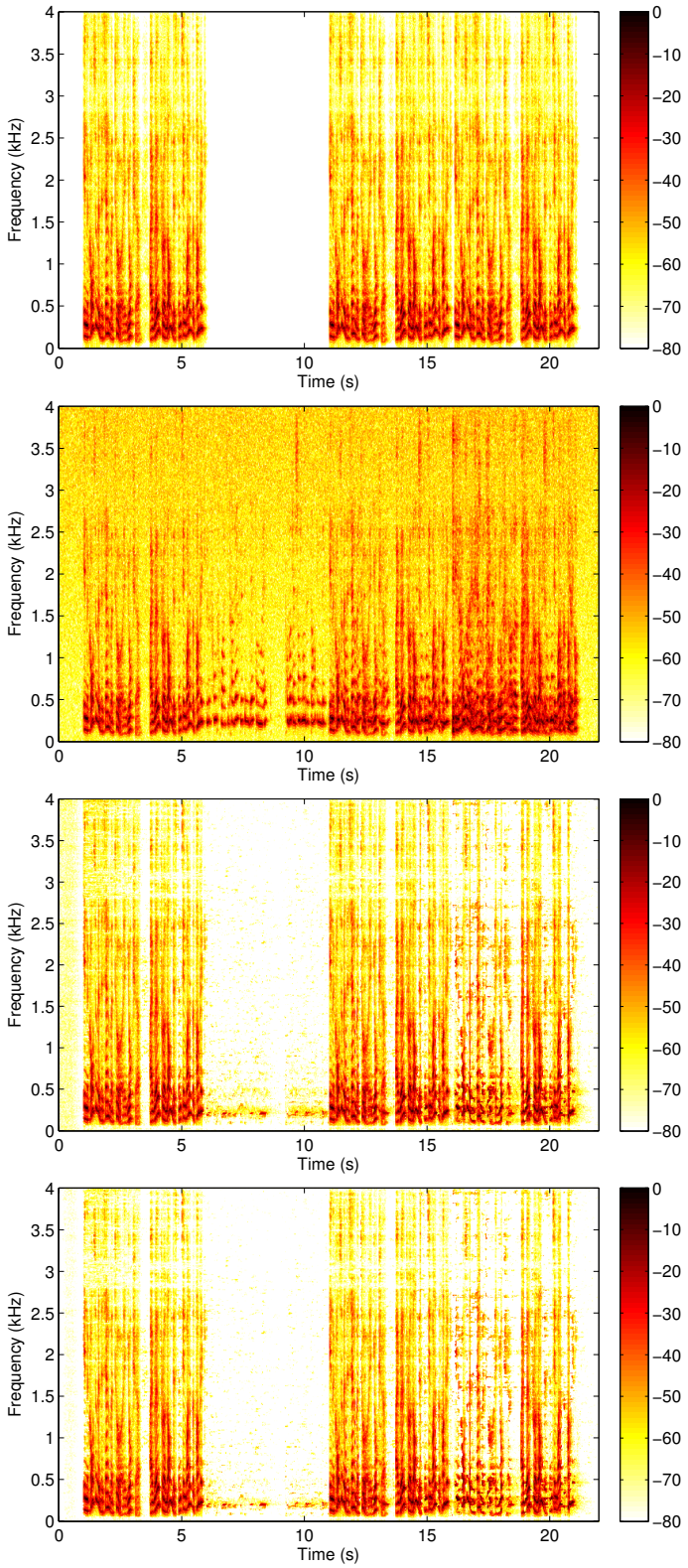


Figure 5.5: Sample spectrograms for an iSCNR of 0 dB: (a) desired speech signal $\widetilde{X}_{00}$, (b) received signal $\widetilde{P}_{00} = \widetilde{X}_{00} + \widetilde{V}_{00,\mathrm{c}} + \widetilde{V}_{00,\mathrm{nc}}$, beamformer output Z for (c) $\eta = 1, \mu' = 0$ and (d) $\eta = 0.25, \mu' = 2$.

5.7 Conclusions

A noise reduction algorithm has been proposed that can distinguish between desired and undesired spatially coherent sources. The desired speech and noise statistics are estimated using a DSPP based on instantaneous high resolution narrowband DOA estimates. The estimated statistics are then applied to a SHD tradeoff beamformer controlled by a tradeoff parameter that can be varied according to the DSPP. A performance evaluation showed that even in the presence of high levels of coherent noise, the proposed algorithm achieved high performance in noise reduction, with SNR improvements of 10–21 dB. These results are in agreement with those of informal listening tests[6].

[6] A number of audio examples are available at
http://www.ee.ic.ac.uk/sap/sphdoa/

Chapter 6

Conclusions

In this chapter, we summarize and conclude the work presented in this thesis. In Sec. 6.1, we highlight its main achievements, and in Sec. 6.2, we outline some suggestions for future research.

6.1 Summary of thesis achievements

The aim of this thesis was to propose a number of acoustic parameter estimation and signal enhancement algorithms for spherical microphone arrays. Its main achievements are as follows:

Acoustic impulse response (AIR) simulation [61,62] In order to comprehensively evaluate the performance of rigid spherical microphone array processing algorithms under a variety of conditions (reverberation time, source-array distance, etc.), it is crucial to be able to simulate their performance in reverberant environments. This is achieved using simulated AIRs, which should incorporate both the effect of the room reflections, and the scattering effects of the rigid sphere. In order to address the lack of AIR algorithms for rigid arrays, in Chapter 3 we proposed such an algorithm based on the combination of a spherical harmonic domain (SHD) scattering model and Allen & Berkley's image method. We also

used the proposed method to investigate the energy of a reverberant sound field, as well as interaural time differences and interaural level differences in both anechoic and echoic environments, based on a rigid spherical head model. Finally, we showed that the proposed method can be used as a mouth simulator, i.e., to simulate the AIR between an omnidirectional microphone and an omnidirectional source positioned on a rigid sphere.

Direction of arrival (DOA) estimation [57] The estimation of the direction of an acoustic source can provide useful information for a number of applications, such as automatic camera steering, noise source identification or beamforming. In Sec. 4.1, we proposed an intensity vector–based DOA estimation approach for spherical microphone arrays, and showed that for a given level of accuracy, it has much lower computational complexity than the steered response power method. Based on simulated AIRs obtained using the algorithm in Chapter 3, we also showed that it is robust to both sensor noise and room reverberation. Finally, an application was presented in Chapter 5, where we used instantaneous DOA estimates to perform noise reduction.

Source tracking [58] In scenarios where the source is moving, DOA estimation becomes more challenging: in order to quickly react to changes in source position, the tracking method must be robust and have low computational complexity. In Sec. 4.2, we proposed a novel source tracking method that meets these requirements, based on an adaptive principal component analysis of the particle velocity vector, which was estimated using the approach presented in Sec. 4.1. It was shown to quickly and accurately track changes in the DOA, even for reverberation times up to 600 ms.

Diffuseness estimation [64] One of the parameters that can be used to describe a sound field is the diffuseness; diffuseness estimates can then be used for dereverberation [18], for example. In Sec. 4.3, we proposed a novel coherence-based

diffuseness estimation method, and compared its performance to a previously proposed spatial domain method, the coefficient of variation method. We showed that for a given amount of time averaging, the estimates obtained using the proposed method have lower variance, even when only zero- and first-order eigenbeams are used.

Noise reduction [56, 59, 63] In distant speech acquisition, noise reduction can be applied to improve the quality and intelligibility of the speech. In Chapter 5, we proposed a tradeoff beamformer in the SHD, which balances the noise reduction performance against speech distortion. The weights of this beamformer depend on the noise and desired signal statistics; accordingly, we proposed a novel statistics estimation algorithm, which can distinguish between desired and undesired spatially coherent sources. The algorithm is based on a desired speech presence probability that is computed based on instantaneous DOA estimates, obtained using the method in Sec. 4.1. We evaluated the complete noise reduction algorithm using measured AIRs, and showed that it achieves high performance, with signal-to-noise ratio improvements of around 10–20 dB.

6.2 Future research directions

In this thesis, we have proposed practical algorithms for acoustic parameter estimation and signal enhancement, evaluated using both simulated and measured impulse responses. Future work should should focus both on improvements to these algorithms, to render them more capable of coping with real acoustic environments and scenarios, and on a more comprehensive evaluation of their performance. Accordingly, we suggest the following future research relating to the work presented in this thesis:

Processing domain: In this work, we performed most processing in the short-time Fourier transform domain with uniform frequency bands. However, perceptually-motivated domains with non-uniform frequency bands have been shown to pro-

vide good subjective performance, e.g., the cepstral domain [19]. As the theory presented is general, the performance of the proposed algorithms could be investigated in other domains.

Acoustic impulse response simulation: The rigid sphere AIR simulation method presented in Chapter 3 could be improved by allowing for *diffuse* room boundary reflections, in addition to specular reflections. Its computational complexity could be reduced by only using the proposed image-based method for the low-order, early reflections, and using a stochastic method to generate the higher-order reflections that make up the reverberant tail of the impulse response, as in [100] where the "diffuse rain" algorithm was used. The accuracy of binaural room impulse responses generated using the proposed method could be improved by using *measured* head-related transfer functions instead of the rigid sphere scattering model.

Direction-of-arrival estimation and tracking: The direction-of-arrival estimation and tracking methods respectively presented in Sec. 4.1 and Sec. 4.2 could be extended to scenarios where *multiple* acoustic sources are present, by working in the time-frequency domain and assuming that only a single source is active in each time-frequency bin (W-disjoint orthogonality [14, 87]).

Diffuseness estimation: Alternative SHD diffuseness estimation methods could be explored, which, while still being based on all the available eigenbeams, would have lower computational complexity.

Noise reduction: The performance of the noise reduction algorithm presented in Chapter 5 could be analyzed in the presence of spatially diffuse noise, as well as, or instead of the spatially incoherent noise used. The performance of the algorithm could also be evaluated in terms of intelligibility using listening tests. In addition, the proposed noise reduction algorithm could be combined with the SHD dereverberation and incoherent noise reduction algorithm presented in [18]

(which uses an estimate of the diffuseness as *a priori* information), to perform joint dereverberation and noise reduction.

Bibliography

[1] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 2002, pp. 1949–1952. ↪ *Cited on pages 27, 47, 102, 117.*

[2] M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York, USA: Dover Publications, 1972. ↪ *Cited on pages 62, 63.*

[3] M. Agmon, B. Rafaely, and J. Tabrikian, "Maximum directivity beamformer for spherical-aperture microphones," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 153–156. ↪ *Cited on page 117.*

[4] J. Ahonen and V. Pulkki, "Diffuseness estimation using temporal variation of intensity vectors," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 285–288. ↪ *Cited on page 110.*

[5] J. Ahonen, V. Pulkki, and T. Lokki, "Teleconference application and B-format microphone array for directional audio coding," in *AES 30th Intl. Conf.*, 2007. ↪ *Cited on page 78.*

[6] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979. ↪ *Cited on pages 49, 50, 51.*

[7] G. B. Arfken and H. J. Weber, *Mathematical Methods for Physicists*, 5th ed. San Diego, CA: Academic Press, 2001. ↪ *Cited on pages 37, 41.*

[8] A. Avni and B. Rafaely, "Sound localization in a sound field represented by spherical harmonics," in *Proc. 2nd Intl. Symp. on Ambisonics and Spherical Acoustics*, Paris, France, 2010, pp. 1–5. ↪ *Cited on page 69.*

[9] I. Balmages and B. Rafaely, "Open-sphere designs for spherical microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 727–732, 2007. ↪ *Cited on page 45.*

[10] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*, ser. SpringerBriefs in Electrical and Computer Engineering. Springer-Verlag, 2011. ↪ *Cited on pages 125, 138.*

[11] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008. ↪ *Cited on pages 26, 117.*

[12] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*. Springer, 2005. ↪ *Cited on pages 124, 125.*

[13] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*. Springer, 2008. No citations.

[14] S. Berge and N. Barrett, "High angular resolution planewave expansion," in *Proc. 2nd Intl. Symp. on Ambisonics and Spherical Acoustics*, 2010. ↪ *Cited on pages 126, 150.*

[15] J. Bermudez, R. C. Chin, P. Davoodian, A. T. Y. Lok, Z. Aliyazicioglu, and H. K. Hwang, "Simulation study on DOA estimation using ESPRIT algorithm," in *Proc. World Congress on Engineering and Computer Science (WCECS)*, vol. 1, 2009, pp. 431–436. ↪ *Cited on page 88.*

[16] T. Betlehem and M. A. Poletti, "Sound field reproduction around a scatterer in reverberation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2009, pp. 89–92. ↪ *Cited on page 50.*

[17] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001. ↪ *Cited on pages 26, 46, 117.*

[18] S. Braun, D. P. Jarrett, J. Fischer, and E. A. P. Habets, "An informed spatial filter for dereverberation in the spherical harmonic domain," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 669–673. ↪ *Cited on pages 148, 150.*

[19] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE Signal Process. Lett.*, vol. 14, no. 12, pp. 1036–1039, Dec. 2007. ↪ *Cited on page 150.*

[20] C. Brown and R. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 476–488, Sep. 1998. ↪ *Cited on page 69.*

[21] Z. Chen and R. C. Maher, "Addressing the discrepancy between measured and modeled impulse responses for small rooms," in *Proc. Audio Eng. Soc. Convention*, Oct. 2007. ↪ *Cited on page 59.*

[22] 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, Sep. 2003. ↪ *Cited on page 118.*

[23] ——, "Multichannel post-filtering in nonstationary noise environments," *IEEE Trans. Signal Process.*, vol. 52, no. 5, pp. 1149–1160, May 2004. ↪ *Cited on page 126.*

[24] I. Cohen, S. Gannot, and B. Berdugo , "An integrated real-time beamforming and post filtering system for nonstationary noise environments," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1064–1073, 2003. ↪ *Cited on page 126.*

[25] R. Compton, Jr., *Adaptive Antennas*, 1st ed. Prentice-Hall, 1988. ↪ *Cited on page 26.*

[26] M. J. Crocker, Ed., *Handbook of Acoustics*. Wiley-Interscience, 1998. ↪ *Cited on pages 84, 93.*

[27] M. J. Crocker and F. Jacobsen, "Sound intensity," in *Handbook of Acoustics*, M. J. Crocker, Ed. Wiley-Interscience, 1998, ch. 106, pp. 1327–1340. ↪ *Cited on pages 84, 131.*

[28] H.-E. de Bree, P. Leussink, T. Korthorst, H. Jansen, T. S. Lammerink, and M. Elwenspoek, "The μ-flown: a novel device for measuring acoustic flows," vol. 54, no. 1-3. Elsevier, 1996, pp. 552–557. ↪ *Cited on pages 84, 96, 131.*

[29] G. Del Galdo, M. Taseska, O. Thiergart, J. Ahonen, and V. Pulkki, "The diffuse sound field in energetic analysis," *J. Acoust. Soc. Am.*, vol. 131, no. 3, pp. 2141–2151, Mar. 2012. ↪ *Cited on page 105.*

[30] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: MacMillan, 1993. ↪ *Cited on page 74.*

[31] S. Doclo and M. Moonen, "On the output SNR of the speech-distortion weighted multichannel Wiener filter," *IEEE Signal Process. Lett.*, vol. 11, no. 12, pp. 809–811, Dec. 2005. ↪ *Cited on page 125.*

[32] J. R. Driscoll and D. M. Healy, "Computing Fourier transforms and convolutions on the 2-sphere," *Advances in Applied Mathematics*, vol. 15, no. 2, pp. 202–250, 1994. ↪ *Cited on pages 40, 41.*

[33] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Am.*, vol. 104, no. 5, pp. 3048–3058, Jan. 1998. ↪ *Cited on pages 55, 58, 69, 71.*

[34] G. W. Elko, "Future directions for microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B.

Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, ch. 17, pp. 383–387. ↪ *Cited on page 26.*

[35] G. W. Elko and J. Meyer, "Spherical microphone arrays for 3D sound recordings," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds., 2004, ch. 3, pp. 67–89. ↪ *Cited on pages 35, 41, 46, 117.*

[36] ——, "Microphone arrays," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer-Verlag, 2008, ch. 50. ↪ *Cited on pages 26, 117.*

[37] European Broadcasting Union. (1988) Sound quality assessment material recordings for subjective tests. http://tech.ebu.ch/publications/sqamcd. [Online]. Available: http://tech.ebu.ch/publications/sqamcd ↪ *Cited on page 135.*

[38] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Analyzing head-related transfer function measurements using surface spherical harmonics," vol. 104, no. 4, pp. 2400–2411, Jun. 1998. ↪ *Cited on page 69.*

[39] R. Fisher, "Dispersion on a sphere," *Proc. Royal Soc. London Ser. A*, vol. 217, no. 1130, pp. 295–305, 1953. ↪ *Cited on page 132.*

[40] R. K. Furness, "Ambisonics - an overview," in *Proc. Audio Eng. Soc. Convention*, Washington, DC, USA, May 1990, pp. 181–189. ↪ *Cited on page 48.*

[41] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001. ↪ *Cited on page 122.*

[42] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer-Verlag, 2008, ch. 47. ↪ *Cited on page 117.*

[43] B. N. Gover, J. G. Ryan, and M. R. Stinson, "Microphone array measurement system for analysis of directional and spatial variations of sound fields," *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 1980–1991, 2002. ↪ *Cited on page 47.*

[44] N. Gumerov and R. Duraiswami, "Modeling the effect of a nearby boundary on the HRTF," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, 2001, pp. 3337–3340. ↪ *Cited on page 50.*

[45] N. A. Gumerov and R. Duraiswami, *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions*. Elsevier, 2005. ↪ *Cited on page 61.*

[46] E. A. P. Habets. (2008, May) Room impulse response (RIR) generator. [Online]. Available: http://home.tiscali.nl/ehabets/rirgenerator.html ↪ *Cited on page 171.*

[47] ——, "A distortionless subband beamformer for noise reduction in reverberant environments," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Aug. 2010, pp. 1–4. ↪ *Cited on page 118.*

[48] E. A. P. Habets and J. Benesty, "Coherent and incoherent interference reduction using a subband tradeoff beamformer," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aug. 2011, pp. 481–485. ↪ *Cited on page 142.*

[49] ——, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 945–958, May 2013. ↪ *Cited on page 138.*

[50] E. A. P. Habets, J. Benesty, and P. A. Naylor, "A speech distortion and interference rejection constraint beamformer," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 854–867, Mar. 2012. ↪ *Cited on pages 117, 122.*

[51] E. A. P. Habets, S. Gannot, and I. Cohen, "Dual-microphone speech dereverberation in a noisy environment," in *Proc. IEEE Intl. Symposium on Signal Processing and Information Technology (ISSPIT)*, Vancouver, Canada, Aug. 2006, pp. 651–655. ↪ *Cited on page 102.*

[52] R. Hendriks and T. Gerkmann, "Noise correlation matrix estimation for multi-microphone speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 223–233, Jan. 2012. ↪ *Cited on page 118.*

[53] ITU-T, *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Mar. 1993. ↪ *Cited on page 136.*

[54] D. P. Jarrett. Spherical Microphone array Impulse Response (SMIR) generator. http://www.ee.ic.ac.uk/sap/smirgen/. [Online]. Available: http://www.ee.ic.ac.uk/sap/smirgen/ ↪ *Cited on pages 50, 61, 64, 88, 98, 137.*

[55] D. P. Jarrett and E. A. P. Habets, "On the noise reduction performance of a spherical harmonic domain tradeoff beamformer," *IEEE Signal Process. Lett.*, vol. 19, no. 11, pp. 773–776, Nov. 2012. ↪ *Cited on pages 104, 122, 123, 136.*

[56] D. P. Jarrett, E. A. P. Habets, J. Benesty, and P. A. Naylor, "A tradeoff beamformer for noise reduction in the spherical harmonic domain," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Aachen, Germany, Sep. 2012. ↪ *Cited on pages 102, 117, 119, 123, 135, 149.*

[57] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 442–446. ↪ *Cited on pages 78, 79, 118, 131, 148.*

[58] ——, "Eigenbeam-based acoustic source tracking in noisy reverberant environments," in *Proc. Asilomar Conf. on Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov. 2010, pp. 576–580. ↪ *Cited on pages 93, 148.*

[59] ——, "Spherical harmonic domain noise reduction using an MVDR beamformer and DOA-based second-order statistics estimation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 654–658. ↪ *Cited on pages 119, 149.*

[60] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, N. D. Gaubitch, and P. A. Naylor, "Dereverberation performance of rigid and open spherical microphone arrays: Theory & simulation," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Edinburgh, UK, Jun. 2011, pp. 145–150. ↪ *Cited on pages 66, 102.*

[61] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Simulating room impulse responses for spherical microphone arrays," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 129–132. ↪ *Cited on pages 50, 147.*

[62] ——, "Rigid sphere room impulse response simulation: algorithm and applications," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1462–1472, Sep. 2012. ↪ *Cited on pages 49, 50, 137, 147.*

[63] D. P. Jarrett, M. Taseska, E. A. P. Habets, and P. A. Naylor, "Noise reduction in the spherical harmonic domain using a tradeoff beamformer and high resolution DOA estimates," *submitted to IEEE Trans. Audio, Speech, Lang. Process.* ↪ *Cited on page 149.*

[64] D. P. Jarrett, O. Thiergart, E. A. P. Habets, and P. A. Naylor, "Coherence-based diffuseness estimation in the spherical harmonic domain," in *Proc. IEEE Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, Eilat, Israel, Nov. 2012. ↪ *Cited on pages 102, 148.*

[65] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2009, pp. 221–224. ↪ *Cited on page 78.*

[66] V. Knudsen and C. Harris, *Acoustical Designing in Architecture*. Wiley, 1950. ↪ *Cited on page 67.*

[67] V. I. Krylov, *Approximate Calculation of Integrals*. New York, NY: MacMillan, 1962. ↪ *Cited on page 41.*

[68] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," *J. Acoust. Soc. Am.*, vol. 62, no. 1, pp. 157–167, 1977. ↪ *Cited on page 69.*

[69] H. Kuttruff, *Room Acoustics*, 4th ed. London: Taylor & Francis, 2000. ↪ *Cited on pages 66, 67, 73, 103, 174.*

[70] E. Lehmann and A. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1429–1439, Aug. 2010. ↪ *Cited on page 49.*

[71] Z. Li and R. Duraiswami, "Flexible and optimal design of spherical microphone arrays for beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 702–714, 2007. ↪ *Cited on page 45.*

[72] ——, "Hemispherical microphone arrays for sound capture and beamforming," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* , 2005, pp. 106–109. ↪ *Cited on pages 45, 58.*

[73] N. A. Logan, "Survey of some early studies of the scattering of plane waves by a sphere," *Proc. of the IEEE*, vol. 53, no. 8, pp. 773–785, Aug. 1965. ↪ *Cited on pages 47, 55.*

[74] K. V. Mardia and P. E. Jupp, *Directional Statistics*. Wiley-Blackwell, 1999. ↪ *Cited on page 132.*

[75] J. Merimaa, "Analysis, synthesis, and perception of spatial sound âĂŞ binaural localization modeling and multichannel loudspeaker reproduction," Ph.D. dissertation, Helsinki University of Technology, 2006. ↪ *Cited on page 85.*

[76] J. Meyer and T. Agnello, "Spherical microphone array for spatial sound recording," in *Proc. Audio Eng. Soc. Convention*, New York, NY, USA, Oct. 2003, pp. 1–9. ↪ *Cited on pages 35, 43, 61.*

[77] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, May 2002, pp. 1781–1784. ↪ *Cited on pages 27, 42, 44, 47, 49, 80, 102.*

[78] J. Meyer and G. W. Elko, "Position independent close-talking microphone," *Signal Processing*, vol. 86, no. 6, pp. 1254–1259, Jun. 2006. ↪ *Cited on pages 38, 56.*

[79] mh acoustics LLC. The Eigenmike microphone array. [Online]. Available: http://www.mhacoustics.com/mh_acoustics/Eigenmike_microphone_array.html ↪ *Cited on pages 40, 43, 58, 135.*

[80] S. Moreau, J. Daniel, and S. Bertet, "3D sound field recording with higher order ambisonics – objective measurements and validation of a 4th order spherical microphone," in *Proc. Audio Eng. Soc. Convention*, May 2006. ↪ *Cited on page 48.*

[81] P. M. Morse and K. U. Ingard, *Theoretical Acoustics*, ser. Intl. Series in Pure and Applied Physics. New York: McGraw Hill, 1968. ↪ *Cited on pages 51, 55.*

[82] A. Nehorai and E. Paldi, "Acoustic vector-sensor array processing," *IEEE Trans. Signal Process.*, vol. 42, no. 9, pp. 2481–2491, Sep. 1994. ↪ *Cited on page 93.*

[83] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. Jensen, "Incorporating the conditional speech presence probability in multi-channel Wiener filter based noise reduction in hearing aids," *EURASIP Journal on Advances in Signal Processing*, vol. Special Issue on Digital Signal Processing for Hearing Instruments, no. 1, 2009. ↪ *Cited on pages 126, 135.*

[84] Y. Peled and B. Rafaely, "Linearly constrained minimum variance method for spherical microphone arrays in a coherent environment," in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, Jun. 2011, pp. 86–91. ↪ *Cited on page 47.*

[85] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Am.*, vol. 80, no. 5, pp. 1527–1529, Nov. 1986. ↪ *Cited on page 49.*

[86] A. D. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*. Acoustical Society of America, 1991. ↪ *Cited on pages 89, 170.*

[87] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, Jun. 2007. ↪ *Cited on pages 102, 126, 150.*

[88] B. D. Radlović, R. Williamson, and R. Kennedy, "Equalization in an acoustic reverberant environment: robustness results," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 311–319, 2000. ↪ *Cited on pages 66, 67, 174.*

[89] B. Rafaely, "Plane-wave decomposition of the pressure on a sphere by spherical convolution," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2149–2157, Oct. 2004. ↪ *Cited on pages 42, 44, 46, 49, 82.*

[90] ——, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005. ↪ *Cited on pages 27, 38, 40, 41, 45, 47, 56, 104, 117, 121, 175.*

[91] ——, "Phase-mode versus delay-and-sum spherical microphone array processing," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 713–716, Oct. 2005. ↪ *Cited on page 46.*

[92] ——, "Spatial sampling and beamforming for spherical microphone arrays," in *Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, May 2008, pp. 5–8. ↪ *Cited on pages 46, 47, 104, 121.*

[93] B. Rafaely and M. Kleider, "Spherical microphone array beam steering using Wigner-D weighting," *IEEE Signal Process. Lett.*, vol. 15, pp. 417–420, 2008. ↪ *Cited on page 45.*

[94] B. Rafaely, B. Weiss, and E. Bachmat, "Spatial aliasing in spherical microphone arrays," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 1003–1010, Mar. 2007. ↪ *Cited on pages 40, 41, 103.*

[95] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, "Spherical microphone array beamforming," in *Speech Processing in Modern Communication: Challenges and Perspectives*, I. Cohen, J. Benesty, and S. Gannot, Eds. Springer, Jan. 2010, ch. 11. ↪ *Cited on pages 38, 46, 81.*

[96] S. Rickard and Z. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Apr. 2002, pp. 529–532. ↪ *Cited on page 126.*

[97] W. W. Rouse Ball, *A Short Account of the History of Mathematics*. Cambridge University Press, 1908. [Online]. Available: http://www.gutenberg.org/ebooks/31246 ↪ *Cited on page 48.*

[98] T. T. Sandel, D. C. Teas, W. E. Feddersen, and L. A. Jeffress, "Localization of sound from single and paired sources," *J. Acoust. Soc. Am.*, vol. 27, no. 5, pp. 842–852, 1955. ↪ *Cited on page 68.*

[99] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964. ↪ *Cited on page 71.*

[100] S. M. Schimmel, M. F. Muller, and N. Dillier, "A fast and accurate "shoebox" room acoustics simulator," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 241–244. ↪ *Cited on page 150.*

[101] D. L. Sengupta, "The sphere," in *Electromagnetic and Acoustic Scattering by Simple Shapes*, J. J. Bowman, T. B. A. Senior, and P. L. E. Uslenghi, Eds. Amsterdam: North-Holland, 1969, ch. 10, pp. 353–415. ↪ *Cited on page 55.*

[102] B. G. Shinn-Cunningham, N. Kopco, and T. J. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *J. Acoust. Soc. Am.*, vol. 117, no. 5, pp. 3100–3115, 2005. ↪ *Cited on pages 69, 73.*

[103] A. Silzle, S. Geyersberger, G. Brohasga, D. Weninger, and M. Leistner, "Vision and technique behind the new studios and listening rooms of the Fraunhofer IIS audio laboratory," in *Proc. Audio Eng. Soc. Convention*, May 2009. ↪ *Cited on page 135.*

[104] 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, Jul. 2010. ↪ *Cited on pages 118, 127, 129, 130, 136.*

[105] ——, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2159–2169, 2011. ↪ *Cited on pages 118, 126, 136.*

[106] SoundField Ltd. SoundField: Benefits of a SoundField system. [Online]. Available: http://www.soundfield.com/soundfield/soundfield.php ↪ *Cited on page 47.*

[107] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004. ↪ *Cited on page 125.*

[108] S. Sra, "A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$," *Computational Statistics*, vol. 27, no. 1, pp. 177–190, 2012. ↪ *Cited on page 132.*

[109] H. Sun, S. Yan, and U. P. Svensson, "Robust minimum sidelobe beamforming for spherical microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 1045–1051, May 2011. ↪ *Cited on page 117.*

[110] F. Talantzis and D. B. Ward, "Robustness of multichannel equalization in an acoustic reverberant environment," *J. Acoust. Soc. Am.*, vol. 114, no. 2, pp. 833–841, 2003. ↪ *Cited on page 67.*

[111] M. Taseska and E. A. P. Habets, "MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sep. 2012. ↪ *Cited on pages 126, 127.*

[112] H. Teutsch, "Wavefield decomposition using microphone arrays and its application to acoustic scene analysis," Ph.D. dissertation, Friedrich-Alexander Universität Erlangen-Nürnberg, 2005. ↪ *Cited on pages 35, 42.*

[113] H. Teutsch and W. Kellermann, "Eigen-beam processing for direction-of-arrival estimation using spherical apertures," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, Piscataway, New Jersey, USA, Mar. 2005. ↪ *Cited on page 78.*

[114] O. Thiergart, G. Del Galdo, and E. A. P. Habets, "On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation," *J. Acoust. Soc. Am.*, vol. 132, no. 4, pp. 2337–2346, 2012. ↪ *Cited on pages 102, 108, 112.*

[115] ——, "Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 309–312. ↪ *Cited on page 102.*

[116] H. L. van Trees, *Detection, Estimation, and Modulation Theory*. New York, USA: Wiley, Apr. 2002, vol. IV, Optimum Array Processing. ↪ *Cited on page 46.*

[117] P. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 511–518. ↪ *Cited on page 118.*

[118] D. B. Ward, "On the performance of acoustic crosstalk cancellation in a reverberant environment," *J. Acoust. Soc. Am.*, vol. 110, pp. 1195–1198, 2001. ↪ *Cited on pages 66, 174.*

[119] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, 1st ed. London: Academic Press, 1999. ↪ *Cited on pages 35, 37, 38, 41, 47, 51, 54, 55, 85, 107, 173, 177.*

[120] S. Yan, H. Sun, U. P. Svensson, X. Ma, and J. M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 361–371, Feb. 2011. ↪ *Cited on pages 47, 117.*

Appendix A

Spherical Microphone array Impulse Response generator (SMIRgen) for MATLAB

An efficient MATLAB/C++ implementation of the rigid sphere acoustic impulse response (AIR) simulation method presented in Chapter 3 has been created. The most computationally complex portions of the method are implemented in the form of a MEX-function. The latest version of the implementation can be downloaded at http://www.ee.ic.ac.uk/sap/smirgen/.

A.1 Documentation

A.1.1 Function call

The function `smir_generator` is called as follows:

```
[h, H, beta_hat] = smir_generator(c, procFs, sphLocation, s, L,
                                  beta, sphType, sphRadius, mic,
                                  N_harm, nsample, K, order);
```

The function input parameters are described below:

Parameter	Description	Default value
`c`	Sound velocity (m/s)	
`procFs`	Sampling frequency (Hz)	
`sphLocation`	Receiver location (x, y, z) in m	
`s`	Source location(s) (x, y, z) in m	
`L`	Room dimensions (x, y, z) in m	
`beta`	Room reflection coefficients $[\beta_{x_1}\ \beta_{x_2}\ \beta_{y_1}\ \beta_{y_2}\ \beta_{z_1}\ \beta_{z_2}]$ or reverberation time T_{60} in s	
`sphType`	Type of sphere (*'open'*/*'rigid'*)	
`sphRadius`	Radius of the sphere (m)	
`mic`	Microphone angles (azimuth, inclination)	
`N_harm`	Maximum order of harmonics to use in spherical harmonic decomposition	
`nsample`	Length of desired AIR	$T_{60} \cdot$ `procFs`
`K`	Oversampling factor	2
`order`	Reflection order (−1 is maximum reflection order)	−1

The function output parameters are described below:

Parameter	Description
`h`	`M` × `nsample` matrix containing the calculated AIR(s)
`H`	`M` × (`K` · `nsample`/2 + 1) matrix containing the calculated ATF(s)
`beta_hat`	If `beta` is the reverberation time, the room reflection coefficient calculated using Sabin-Franklin's formula [86] is returned.

A.1.2 Notes

- The most computationally complex parts of this algorithm have been placed in a C++ function with a MEX wrapper. To use it you will need to build the MEX-function using MATLAB's `mex` command.

- The functions `mysph2cart()` and `mycart2sph()` are included in order to convert between spherical and Cartesian coordinates. These functions use the coordinate systems defined in Sec. 2.1. The microphone angles used as inputs to SMIRgen must be obtained using `mycart2sph()`, or use the same coordinate system. Specifically, azimuth is measured counterclockwise from the positive x axis (the positive y axis has an azimuth of 90°) and inclination is measured from the positive z axis (the x-y plane has an inclination of 90°).

- When the source-array distance is small, it is necessary to oversample in the frequency domain in order to avoid the wrap-around effect of the discrete Fourier transform. For this purpose, choose $K > 1$, e.g., $K = 2$ or $K = 4$.

- The example script `run_smir_generator_comparison` compares the output of SMIRgen to the output of Emanuël Habets's RIR generator [46], with each of the array's microphones treated as a separate receiver. This comparison is only valid in the open sphere case, since the RIR generator does not account for scattering. A copy of the AIR generator is included for this purpose, in accordance with the terms of the GNU General Public License. To use it you will need to build the MEX-function using MATLAB's `mex` command.

A.2 Example

An example of an AIR and acoustic transfer function (ATF) generated using SMIRgen is provided in Fig. A.1. The following input parameters were used:

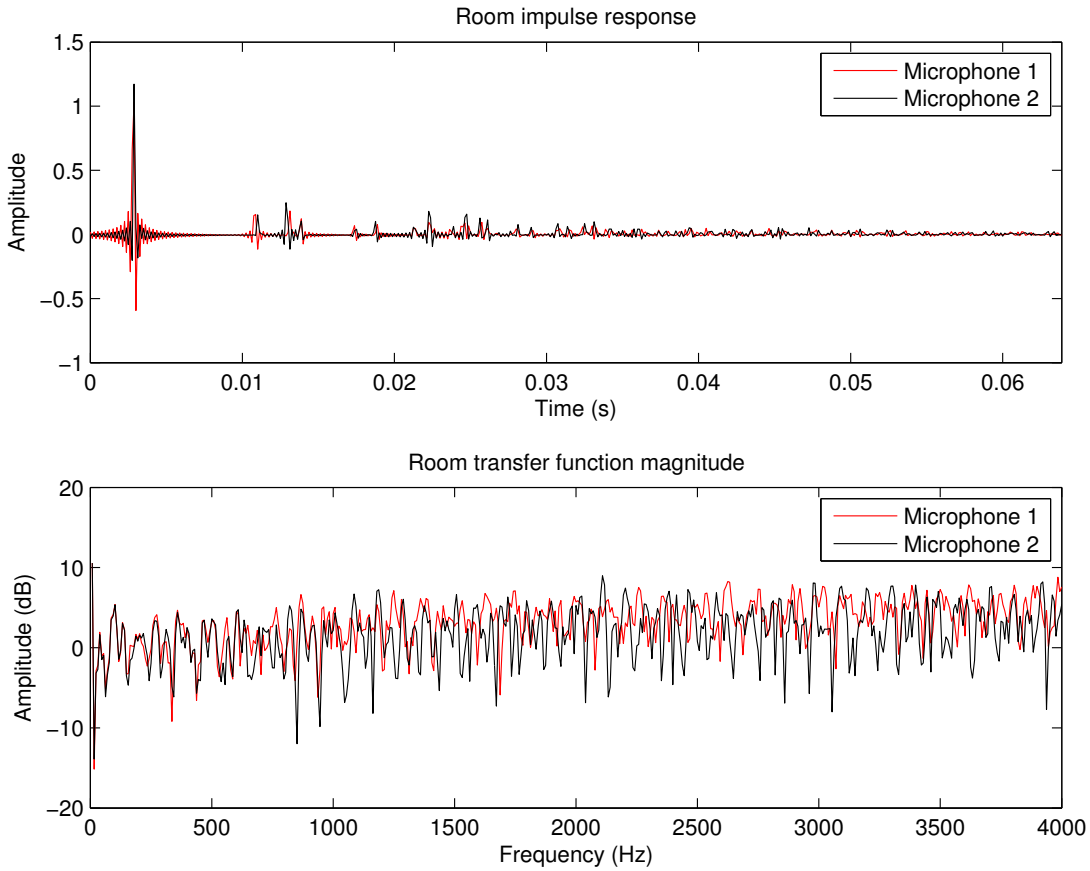


Figure A.1: Sample acoustic impulse response and acoustic transfer function generated using SMIRgen.

```
procFs = 8000;
c = 343;
nsample = 512;
N_harm = 40;
K = 2;
L = [4 6 8];
sphLocation = [2 3.2 4];
s = [2.37 4.05 4.4];
beta = [1 0.7 0.7 0.5 0.2 1];
order = -1;
sphRadius = 0.042;
sphType = 'rigid';
mic = [pi/4 pi/2; 3*pi/4 pi/2];
```

Appendix B

Spatial correlation in a diffuse sound field

The sound pressure at a position $\tilde{\mathbf{r}} = (r, \Omega)$ due to a unit amplitude plane wave incident from direction Ω_0 is given by [119]

$$P(\tilde{\mathbf{r}}, \Omega_0, k) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} 4\pi\varphi(\Omega_0) b_l(k) Y_{lm}^*(\Omega_0) Y_{lm}(\Omega), \tag{B.1}$$

where $\varphi(\Omega_0)$ is a random phase term and $|\varphi(\Omega_0)| = 1$. Assuming a diffuse sound field, the spatial cross-correlation between the sound pressure at two positions $\tilde{\mathbf{r}} = (r, \Omega)$ and $\tilde{\mathbf{r}}' = (r, \Omega')$ is given by:

$$\begin{aligned} C(\tilde{\mathbf{r}}, \tilde{\mathbf{r}}', k) =& \frac{1}{4\pi} \int_{\Omega_0 \in \mathcal{S}^2} P(\tilde{\mathbf{r}}, \Omega_0, k) P^*(\tilde{\mathbf{r}}', \Omega_0, k) \mathrm{d}\Omega_0 \\ =& \frac{1}{4\pi} \int_{\Omega_0 \in \mathcal{S}^2} \sum_{l=0}^{\infty} \sum_{m=-l}^{l} 4\pi b_l(k) Y_{lm}^*(\Omega_0) Y_{lm}(\Omega) \\ & \sum_{l'=0}^{\infty} \sum_{m'=-l'}^{l'} 4\pi b_{l'}^*(kr) Y_{l'm'}(\Omega_0) Y_{l'm'}^*(\Omega') \mathrm{d}\Omega_0. \end{aligned}$$

Using the orthonormality property of the spherical harmonics in (2.6) and the addition theorem in (3.8), we eliminate the cross terms followed by the sum over m and obtain

$$C(\tilde{\mathbf{r}}, \tilde{\mathbf{r}}', k) = \frac{1}{4\pi} \sum_{l=0}^{\infty} \sum_{m=-l}^{l} (4\pi)^2 |b_l(k)|^2 Y_{lm}(\Omega) Y_{lm}^*(\Omega') \tag{B.2a}$$

$$= \frac{1}{4\pi} \sum_{l=0}^{\infty} (4\pi)^2 |b_l(k)|^2 \frac{2l+1}{4\pi} \mathcal{P}_l(\cos \Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{r}}'}) \tag{B.2b}$$

$$= \sum_{l=0}^{\infty} |b_l(k)|^2 (2l+1) \mathcal{P}_l(\cos \Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{r}}'}), \tag{B.2c}$$

where $\Theta_{\tilde{\mathbf{r}},\tilde{\mathbf{r}}'}$ is the angle between $\tilde{\mathbf{r}}$ and $\tilde{\mathbf{r}}'$.

In the open sphere case where $b_l(k) = (-i)^l j_l(kr)$, we can express (B.2a) as

$$C(\tilde{\mathbf{r}}, \tilde{\mathbf{r}}', k) = \Im \left\{ 4\pi i \sum_{l=0}^{\infty} \sum_{m=-l}^{l} |b_l(k)|^2 Y_{lm}(\Omega) Y_{lm}^*(\Omega') \right\}$$

$$= \Im \left\{ 4\pi i \sum_{l=0}^{\infty} \sum_{m=-l}^{l} j_l(kr) h_l^{(1)}(kr) Y_{lm}(\Omega) Y_{lm}^*(\Omega') \right\}$$

using $\Re\{h_l^{(1)}(kr)\} = j_l(kr)$, where $\Re$ and $\Im$ respectively denote the real and imaginary parts of a complex number. Finally, using (3.7), we obtain the well-known spatial domain result for two omnidirectional receivers in a diffuse sound field [69, 88, 118]:

$$C(\tilde{\mathbf{r}}, \tilde{\mathbf{r}}', k) = \Im \left\{ \frac{e^{+ik\|\tilde{\mathbf{r}}-\tilde{\mathbf{r}}'\|}}{k\|\tilde{\mathbf{r}}-\tilde{\mathbf{r}}'\|} \right\}$$

$$= \frac{\sin(k\|\tilde{\mathbf{r}}-\tilde{\mathbf{r}}'\|)}{k\|\tilde{\mathbf{r}}-\tilde{\mathbf{r}}'\|}. \tag{B.3}$$

Appendix C

Relationship between the zero-order eigenbeam and the omnidirectional reference microphone signal

Property C.0.1. Let $P_{lm}(k)$ denote the spherical Fourier transform, as defined in (2.4), of the spatial domain sound pressure $P(k, \mathbf{r})$, where $\mathbf{r}$ denotes the position (in spherical coordinates) with respect to the centre of a spherical microphone array with mode strength $b_l(k)$. Let $P_{\mathcal{M}_{\text{ref}}}(k)$ denote the sound pressure which would be measured, were an omnidirectional microphone $\mathcal{M}_{\text{ref}}$ to be placed at a position corresponding to the centre of the sphere, i.e., at the origin of the spherical coordinate system; $P_{\mathcal{M}_{\text{ref}}}(k)$ is then related to the zero-order eigenbeam $P_{00}(k)$ via the relationship[1]

$$P_{\mathcal{M}_{\text{ref}}}(k) = \frac{P_{00}(k)}{\sqrt{4\pi}\, b_0(k)}. \tag{C.1}$$

[1]It should be noted that this relationship is dependent upon the chosen mode strength definition (see Sec. 2.4). If a 4π factor is included in $b_l(k)$, as in [90], the relationship becomes $P_{\mathcal{M}_{\text{ref}}}(k) = \sqrt{4\pi}\frac{P_{00}(k)}{b_0(k)}$.

Proof. We assume, without loss of generality[2], that the sound field is composed of a single spherical wave incident from a point source at a position $\mathbf{r}_s = (r_s, \Omega_s)$, in which case the spatial domain sound pressure $P(k, \mathbf{r})$ is given by (3.11), i.e.,

$$P(k, \mathbf{r}) = k \sum_{l=0}^{\infty} (-i)^{-(l+1)} b_l(k) h_l^{(1)}(kr_s) \sum_{m=-l}^{l} Y_{lm}^*(\Omega_s) Y_{lm}(\Omega). \tag{C.2}$$

From the definition of the spherical Fourier transform, $P_{00}(k)$ is given by

$$P_{00}(k) = \int_{\Omega \in \mathcal{S}^2} P(k, \mathbf{r}) Y_{00}^*(\Omega) \mathrm{d}\Omega. \tag{C.3}$$

By substituting (C.2) into (C.3), we find

$$P_{00}(k) = \int_{\Omega \in \mathcal{S}^2} k \sum_{l=0}^{\infty} (-i)^{-(l+1)} b_l(k) h_l^{(1)}(kr_s) \sum_{m=-l}^{l} Y_{lm}^*(\Omega_s) Y_{lm}(\Omega) Y_{00}^*(\Omega) \mathrm{d}\Omega. \tag{C.4}$$

Using the orthonormality of the spherical harmonics (2.6) and the fact that $Y_{00}(\cdot) = 1/\sqrt{4\pi}$, we can simplify (C.4) to

$$P_{00}(k) = k(-i)^{-1} b_0(k) h_0^{(1)}(kr_s) Y_{00}^*(\Omega_s) \tag{C.5a}$$

$$= \frac{ik}{\sqrt{4\pi}} b_0(k) h_0^{(1)}(kr_s). \tag{C.5b}$$

Furthermore, in the absence of the sphere, the sound pressure measured at a point $\mathbf{r} = \mathbf{0}$ due to a single spherical wave incident from a point source at a position $\mathbf{r}_s = (r_s, \Omega_s)$ is given by (3.7), i.e.,

$$P_{\mathcal{M}_{\mathrm{ref}}}(k) = \frac{e^{ik\|\mathbf{r}_s\|}}{4\pi \|r_s\|} \tag{C.6a}$$

$$= \frac{e^{ikr_s}}{4\pi r_s}. \tag{C.6b}$$

[2]The operations involved in the proof are linear, and the proof therefore holds for any number of spherical waves.

Finally, using the fact that $h_0^{(1)}(x) = \frac{e^{ix}}{ix}$ [119, eqn. 6.62], we can simplify (C.5b) to

$$P_{00}(k) = \frac{ik}{\sqrt{4\pi}} b_0(k) \frac{e^{ikr_s}}{ikr_s} \tag{C.7a}$$

$$= \sqrt{4\pi} b_0(k) \frac{e^{ikr_s}}{4\pi r_s} \tag{C.7b}$$

$$= \sqrt{4\pi} b_0(k) P_{\mathcal{M}_{\text{ref}}}(k), \tag{C.7c}$$

and therefore Property C.0.1 holds. □