

Department of Signal Processing and Acoustics

Parametric spatial audio processing utilising compact microphone arrays

Symeon Delikaris-Manias



Parametric spatial audio processing utilising compact microphone arrays

Symeon Delikaris-Manias

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the Auditorium F239a, Otakaari 3, Espoo on November 10, 2017 at 12 noon.

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics
Communication Acoustics Group

Supervising professor

Professor Ville Pulkki

Preliminary examiners

Professor Ivan Tashev, Microsoft Research, USA

Doctor Mehrez Souden, Apple Inc., USA

Opponent

Professor Craig Jin, University of Sydney, Australia

Aalto University publication series

DOCTORAL DISSERTATIONS 197/2017

© 2017 Symeon Delikaris-Manias

ISBN 978-952-60-7661-4 (printed)

ISBN 978-952-60-7660-7 (pdf)

ISSN-L 1799-4934

ISSN 1799-4934 (printed)

ISSN 1799-4942 (pdf)

<http://urn.fi/URN:ISBN:978-952-60-7660-7>

Images: Symeon Delikaris-Manias (unless otherwise stated). Photo on the cover: lake Kämmenlampi where a nearly-perfect representation of the forest is visible through its reflection

Unigrafia Oy
Helsinki 2017

Finland

Author

Symeon Delikaris-Manias

Name of the doctoral dissertation

Parametric spatial audio processing utilising compact microphone arrays

Publisher School of Electrical Engineering**Unit** Department of Signal Processing and Acoustics**Series** Aalto University publication series DOCTORAL DISSERTATIONS 197/2017**Field of research** Acoustics and signal processing**Manuscript submitted** 12 June 2017**Date of the defence** 10 November 2017**Permission to publish granted (date)** 16 August 2017**Language** English **Monograph** **Article dissertation** **Essay dissertation****Abstract**

This dissertation focuses on the development of novel parametric spatial audio techniques using compact microphone arrays. Compact arrays are of special interest since they can be adapted to fit in portable devices, opening the possibility of exploiting the potential of immersive spatial audio algorithms in our daily lives. The techniques developed in this thesis consider the use of signal processing algorithms adapted for human listeners, thus exploiting the capabilities and limitations of human spatial hearing. The findings of this research are in the following three areas of spatial audio processing: directional filtering, spatial audio reproduction, and direction of arrival estimation.

In directional filtering, two novel algorithms have been developed based on the cross-pattern coherence (CroPaC). The method essentially exploits the directional response of two different types of beamformers by using their cross-spectrum to estimate a soft masker. The soft masker provides a probability-like parameter that indicates whether there is sound present in specific directions. It is then used as a post-filter to provide further suppression of directionally distributed noise at the output of a beamformer. The performance of these algorithms represent a significant improvement over previous state-of-the-art methods.

In parametric spatial audio reproduction, an algorithm is developed for multi-channel loudspeaker and headphone rendering. Current limitations in spatial audio reproduction are related to high inter-channel coherence between the channels, which is common in signal-independent systems, or time-frequency artefacts in parametric systems. The developed algorithm focuses on solving these limitations by utilising two sets of beamformers. The first set of beamformers, namely analysis beamformers, is used to estimate a set of perceptually-relevant sound-field parameters, such as the separate channel energies, inter-channel time differences and inter-channel coherences of the target-output-setup signals. The directionality of the analysis beamformers is defined so that it follows that of typical loudspeaker panning functions and, for headphone reproduction, that of the head-related transfer functions (HRTFs). The directionality of the second set of high audio quality beamformers is then enhanced with the parametric information derived from the analysis beamformers. Listening tests confirm the perceptual benefit of such type of processing.

In direction of arrival (DOA) estimation, histogram analysis of beamforming and active intensity based DOA estimators has been proposed. Numerical simulations and experiments with prototype and commercial microphone arrays show that the accuracy of DOA estimation is improved.

Keywords spatial audio, directional filtering, perceptual sound reproduction, microphone arrays**ISBN (printed)** 978-952-60-7661-4**ISBN (pdf)** 978-952-60-7660-7**ISSN-L** 1799-4934**ISSN (printed)** 1799-4934**ISSN (pdf)** 1799-4942**Location of publisher** Helsinki**Location of printing** Helsinki**Year** 2017**Pages** 160**urn** <http://urn.fi/URN:ISBN:978-952-60-7660-7>

Preface

This work was carried out at the Department of Signal Processing and Acoustics, School of Electrical Engineering, Aalto University in Espoo, Finland. The research leading to these results has received funding from the European Research Council under the European Community Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no 240453 and personal scholarships from the Aalto ELEC Doctoral school, the Aalto Foundation and the Nokia Foundation.

My gratitude goes to Professori Pulkki. The research has been very pleasant under his supervision but also in the cottage while working near nature. I am grateful for the freedom I have been given for scientific research but at the same for the guidance whenever I needed it.

Special thanks to my co-authors for the very enjoyable collaborations: Dr. Juha Vilkamo for the endless brainstorming and back-and-forth-style paper writing, Leo McCormack for his enthusiasm in real-time implementations, Despoina Pavlidi for the fruitful scientific discussions and Prof. Athanasios Mouchtaris for his insightful commenting during the compilation of our manuscripts. I'm also greatly appreciative to the pre-examiners Prof. Ivan Tashev and Dr. Mehrez Souden, and Prof. Craig Jin for accepting to be my opponent.

The current and former people at the Acoustics Lab have made my stay incredible fun: group meetings in the mökki, paddling in the archipelago of Helsinki, band rehearsals and gigs, synthesiser meetings and futsal games at the piano bar. Many thanks goes to Dr. Stefano D' Angelo, Dr.

Jukka Ahonen, Allesandro Altoé, Fabian Esqueda Flores, Dr. Catarina Hiipakka, Dr. Marko Hiipakka, Dr. Sofoklis Kakouros, Teemu Koski, Dr. Mikko-Ville Laitinen, Oliver Merilaid, Juhani Paasonen, Tapani Pihlajakuja, Dr. Jouni Pohjalainen, Dr. Archontis Politis, Henri Pöntynen Dr. Tuomo Raitio, Dr. Olli Rummukainen, Ville Saari, Dr. Olli Santala, Dr. Marko Takanen, Julia Turku and all the wonderful people from the Audio Signal Processing, Speech Technology and Virtual Acoustics group. Special thanks goes to Ilkka Huhtakallio for sharing his knowledge and enthusiasm in acoustics and audio systems and for his support.

Sincere thanks to Prof. Boaz Rafaely for hosting me during my research visit and all the researchers of the group who have made my stay at Beer-sheba and Tel Aviv incredibly nice: Dr. Hai Morgenstern, Dr. Noam Shab-tai, Yoav Biderman, Uri Abend, Zamir Ben-Hur, Amir Musicant and to Dr. David Alon for the very pleasant collaboration. I really enjoyed the relaxed and at the same time educational and inspiring working atmosphere. To my colleagues at Apple where I had the chance to work with many talented people, thanks for making my stay very enjoyable and special thanks to Izzy for introducing me to some beautiful places in SoCal. I am also grateful to the following hosts around the world and specifically to Dr. Tad Rollow, Prof. Julius O. Smith, Prof. Thushara Abhayapala, Dr. Joshua Reiss, Prof. Tuomas Virtanen and Prof. Samuli Siltanen.

Most importantly of all I would like to express my gratitude to my family and especially to my mother for showing me how to enjoy life and go after what I take pleasure most.

Helsinki, October 9, 2017,

Symeon Delikaris-Manias

Contents

Preface	1
Contents	3
List of Publications	7
Author's Contribution	9
List of Figures	13
1. Introduction	17
2. Background	19
2.1 Physical properties of a sound field	19
2.2 Capturing the properties of a sound field	21
2.3 Microphone arrays	22
3. Perception	25
3.1 Human auditory system	25
3.2 Time-frequency resolution of the human auditory system . .	27
3.3 Sound source localisation	29
3.3.1 Inter-aural time and level differences	29
3.3.2 Head-related transfer function	29
3.3.3 Dynamic cues	30
3.3.4 Precedence effect	30
3.3.5 Perception of distance	30
3.4 Time-frequency resolution for perceptually-motivated sig- nal processing	31

4. Directional filtering and analysis of spatial sound	33
4.1 Signal-independent directional noise reduction	34
4.2 Signal-dependent directional noise reduction	36
4.2.1 Beamforming	36
4.2.2 Post filtering	37
4.3 Spherical harmonic analysis	41
4.4 Direction-of-arrival estimation	43
4.5 Acoustic camera for sound field visualisation	44
5. Reproduction of spatial sound	47
5.1 Overview	47
5.2 Signal-independent reproduction	47
5.2.1 Channel-based systems	48
5.2.2 Object or panning-based systems	50
5.2.3 Scene-based systems	51
5.3 Parametric reproduction	52
5.3.1 Inter-channel level difference, time difference and co- herence	54
5.3.2 Direction of arrival and sparsity assumption	55
5.3.3 Direction of arrival and diffuseness	55
5.3.4 Target-setup parametrisation with optimal mixing	56
5.3.5 Object-based parametrisation	57
6. Summary of contributions	59
6.1 Directional filtering utilising the cross-pattern coherence	59
6.2 Directional filtering based on orthogonally-weighted beam- formers in the spherical harmonic domain	60
6.3 Perceptually-motivated spatial sound reproduction based on optimal mixing	60
6.4 Parametric binaural reproduction for compact microphone arrays	61
6.5 Direction-of-arrival estimation with histogram analysis of steered-response beamformers	62
6.6 Direction-of-arrival estimation with histogram analysis of spatially constrained active intensity vectors	62
6.7 Acoustic camera based on spatial parameters	63

7. Conclusions	65
References	67
Publications	79

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Symeon Delikaris-Manias and Ville Pulkki. Cross pattern coherence algorithm for spatial filtering applications utilizing microphone arrays. *IEEE Transactions on Audio, Speech, and Language Processing*, Volume 21, issue 11, pages 2356–2367, November 2013.

II Symeon Delikaris-Manias, Juha Vilkkamo, and Ville Pulkki. Signal-dependent spatial filtering based on weighted-orthogonal beamformers in the spherical harmonic domain. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, Volume 24, issue 9, pages 1507–1519, April 2016.

III Juha Vilkkamo and Symeon Delikaris-Manias. Perceptual reproduction of spatial sound using loudspeaker-signal-domain parametrization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Volume 23, issue 10, pages 1660–1669, June 2015.

IV Symeon Delikaris-Manias, Juha Vilkkamo, and Ville Pulkki. Parametric binaural rendering utilizing compact microphone arrays. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, pages 629–633, 19–24 April 2015.

- V** Symeon Delikaris-Manias, Despoina Pavlidi, Ville Pulkki, and Athanasios Mouchtaris. 3D localization of multiple audio sources utilizing 2D DOA histograms. In *24th European Signal Processing Conference (EU-SIPCO)*, Budapest, Hungary, pages 1473–1477, 29 August–2 September 2016.
- VI** Symeon Delikaris-Manias, Despoina Pavlidi, Athanasios Mouchtaris, and Ville Pulkki. DOA estimation with histogram analysis of spatially constrained intensity vectors. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, pages 526–530, 5–9 March 2017.
- VII** Leo McCormack, Symeon Delikaris-Manias, and Ville Pulkki. Parametric acoustic camera for real-time sound capture, analysis and tracking. In *International Conference on Digital Audio Effects (DAFx-17)*, Edinburgh, UK, 5–9 September 2017.

Author's Contribution

Publication I: “Cross pattern coherence algorithm for spatial filtering applications utilizing microphone arrays”

The main idea was invented by the co-author and was formulated and developed by the present author. The present author designed the experiments with the co-author and was responsible for conducting the experiments and writing the article. The present author wrote the article with comments by the co-author.

Publication II: “Signal-dependent spatial filtering based on weighted-orthogonal beamformers in the spherical harmonic domain”

The idea of using coherence-based measures between two types of beamformers was invented by the present author. The algorithm was developed together with the second author. The mathematical proofs were derived by the present author. The present author wrote the article with comments by the second author. The second author developed the perceptually-motivated filter bank that was utilised in the study.

Publication III: “Perceptual reproduction of spatial sound using loudspeaker-signal-domain parametrization”

The idea of estimating perceptually-motivated spatial parameters using a set of constant directivity analysis beamformers was invented by the

present author together with the co-author as a result of a discussion how on to solve issues related to spatial sound reproduction artefacts using non-parametric and state-of-the-art parametric methods. The development, evaluation, and analysis of the article was shared between the present and the second author.

Publication IV: “Parametric binaural rendering utilising compact microphone arrays”

The idea of adaptive mixing between two types of beamformers for binaural reproduction was suggested by the present author. This idea extends the sound-field parametrisation algorithm that was developed for Publication III for headphone reproduction. The development, analysis and writing of the paper was shared between the present and the second author.

Publication V: “3D localization of multiple audio sources utilizing 2D DOA histograms”

The idea to use histogram analysis in existing scanning-based DOA estimators was suggested by the present author. The present and second author developed the DOA estimator. The evaluation was conducted together with the second author. The present author was mainly responsible of writing the article with comments by the second author.

Publication VI: “DOA estimation with histogram analysis of spatially constrained intensity vectors”

The present author investigated the use of the spatially constrained intensity vectors as DOA estimators. The present author developed and implemented the DOA estimator. The second author developed the post-processing of the DOA estimator. The evaluation was conducted together with the second author. The present author was mainly responsible of writing the article with comments by the second author.

Publication VII: “Parametric acoustic camera for real-time sound capture, analysis and tracking”

The present author investigated the use of the cross-pattern coherence algorithm for sound-field visualisation. The initial idea of the side-lobe suppression algorithm was invented by the present author. The formulation of the algorithm was performed by the present and first author. The experiments were conducted together with the first author. The development of the plugins were performed by the first author in collaboration with the present author. The writing of the article was shared between the present and the first author.

Author's Contribution

List of Figures

2.1	Impulse response of the a concert hall (top) and the energy decay (bottom).	20
2.2	Frequency response over time of the RIR depicted in Fig. 2.1.	21
2.3	Polar graphs on a logarithmic scale of ideal first-order microphone.	22
2.4	Prototype compact microphone arrays constructed for experiments during the dissertation. From left to right: open 5-channel circular microphone array or 1.5 cm radius, different-radii wooden rigid cylinders able to fit 5 to 8 microphones, 8-channel cylinder of 1.3 cm radius, 8-channel cylinder or 4 cm radius, 3-channel mobile-like array made with styro-foam and 7-channel mobile-like array made out of wood. . .	23
2.5	Angular frequency response of a single omnidirectional microphone fitted on the surface of a rigid mobile-like device. .	24
3.1	Human auditory system [1].	26
3.2	Sound arriving from a single source (norsu) to a listener. Human hearing exploits simultaneously different cues that contribute to determining the location of sounds. Sounds can arrive at the two ears with different delays and or different levels. The head shadowing and reflection from the pinna and torso provide spectra-temporal information. . . .	28
4.1	An illustrative scenario where multiple sources are captured with a microphone array. The task is to enhance sounds originating from the desired region (indicated with the darker triangle).	34

4.2	Beamforming: a signal-independent approach (left), where the signals are simply mixed with time-invariant weights $w_{ds/ls}(k)$, and a signal-dependent approach (right) where sub-band weights $w_a(k, n)$ for time index n are applied to a time-frequency (TF) transformed signal from the microphones. . .	35
4.3	Post-filtering a beamformer	38
4.4	Illustration of the resulting directional selectivity of a post-filter derived by multiplying the output of two types of beamformers. Note that the beamformers have the same phase and equal magnitude in the same direction. The half-wave rectification process, shown as the maximum operation, ensures that any directional information arriving with negative phase is neglected. Adopted from [2].	40
4.5	Block diagram for DOA estimation using the active intensity.	44
5.1	Generalised components of spatial sound reproduction: capture, transmission and rendering.	48
5.2	State-of-the-art signal-independent spatial sound reproduction systems: channel-based (left), object-based (middle) and scene-based (right).	49
5.3	Parametric spatial-sound reproduction. Time-frequency representations of microphone array signals are analysed in the capturing stage, the microphone signals or a down-mixed signal of them is transmitter along with a set of parameters. At the rendering stage the captured sound scene is synthesised for an arbitrary loudspeaker or headphone setup. . . .	53

List of Abbreviations

BCC	binaural cue coding
BSS	blind source separation
CroPaC	cross-pattern coherence
DaS	delay and sum
DirAC	directional audio coding
DOA	direction of arrival
FOA	first-order ambisonics
HOA	higher-order ambisonics
HRTF	head-related transfer function
ICC	inter-channel coherence
ICLD	inter-channel level difference
ICTD	inter-channel time difference
ILD	inter-aural level difference
ITD	inter-aural time difference
LCMV	linearly-constrained minimum variance
MUSIC	multiple signal classification
MVDR	minimum-variance distortionless response
SMA	spherical microphone array

List of Symbols

α	average absorption coefficient
RT_{60}	reverberation time
w_{PWD}	spherical harmonic beamforming weights
C_{lm}	covariance matrix of the spherical harmonic signals
d	axis-symmetric spherical harmonic beamforming coefficients
I	instantaneous intensity
s	vector of spherical harmonic signals
W	spatial encoding matrix
w_a	beamforming weights for adaptive beamformers
$w_{\text{ds/ls}}$	beamforming weights for delay and sum or least squares methods
x	microphone array input signals
A	absorption area
f_c	centre frequency of the equivalent rectangular bandwidth filters
f_{ERB}	width of the equivalent rectangular bandwidth filters
G	post filter
L	order of the spherical harmonic signals
n	time index
p	pressure
Q	number of microphones
S	surface area
u	particle velocity
u_x	angular response of pressure-gradient microphone
V	volume of the room
w_p	weighting factor for the polar equation

1. Introduction

Spatial hearing provides us with information about our surroundings: the location of multiple sound sources along with their corresponding distance, information about the space they reside in and their content. Parametric spatial sound technologies are multichannel signal processing algorithms that aim to capture, analyse and render a complex sound scene for a human listener in order that he or she obtains an understanding of the environment or has the auditory experience of being there.

The topic of the dissertation is the development of parametric spatial sound algorithms using compact microphone arrays. Compact microphone arrays are especially appealing; they are portable and can be easily fitted in mobile or handheld devices thus extending the applicability of microphone array-based spatial sound signal processing algorithms to devices being utilised in everyday life situations, sometimes even as personal assistants. A sound-focusing algorithm can, for example, extend our abilities for a clearer understanding of a complex sound environment due to its ability to obtain information only from specific directions within a sound scene while discarding all other sound information. This concept can be then extended to provide an enhanced understanding of the whole sound scene that was not previously possible. An application of such algorithms is the reconstruction and/or modification of a sound scene that can give the impression of being there. Improving, enhancing, and providing such new experiences can improve some aspects of the quality of life.

This dissertation considers the concepts of capturing, manipulating, and rendering a sound scene with compact microphone arrays. Capturing a sound scene consists of utilising multiple microphones to record and analyse its characteristics. The manipulation of a sound scene is defined by the

target application whether this involves the analysis of specific directions or the whole 3-D space. The rendering part then collects all information from the capturing and manipulation stages and visualises it or renders it over loudspeakers or headphones.

The term parametric refers to the fact that in the capturing stage a single or multiple parameters able to describe the sound scene are estimated. Two application classes of spatial audio processing are considered within the context of this thesis: directional filtering and reproduction. Both classes are perceptually motivated, meaning that signal processing techniques are adapted for human listeners, exploiting their capabilities and limitations. The improvement of conventional signal processing algorithms that can estimate parameters such as direction of arrival (DOA) using the spatially constrained active intensity and steered-response power beamformers are also studied in this thesis, due to their usefulness within the context of directional filtering. DOA estimators can provide information about the sound sources within an environment and their trajectories. A directional filtering algorithm can then utilise this information for focusing in single or multiple directions of interest.

The algorithms developed in this thesis were evaluated using numerical simulations and/or prototype devices that were specifically built to show a more realistic performance. Prototype microphone arrays as well as commercial ones were utilised to capture the sound field. In those cases where the spatial audio algorithm implemented in this thesis involved rendering through headphones or loudspeakers, the whole system has been evaluated not only with instrumental measures but also with listening tests.

This introduction provides an overview of the basic concepts of sound, spatial hearing, and spatial sound processing technologies. The main concepts are explained here while the technical details can be found in the published articles. The introduction is organised as follows: Section 2 provides an overview of spatial sound as a physical phenomenon and how to capture it with microphones. Section 3 provides an overview of the human perception of spatial sound, section 4 deals with the directional filtering as a spatial sound technology, section 5 provides an overview of spatial sound reproduction technologies, section 6 summarises the main contributions, and section 7 provides the conclusions.

2. Background

Three-dimensional sound scenes are part of our everyday life experiences: having a conversation in a room, walking in a crowded street, or listening to an orchestra in a concert hall. These sound scenes consist of single or multiple sound sources and their interaction with the space. A complex sound scene can be seen as a superposition of sound waves. A sound wave is produced when particles in a medium undergo vibrations. Such waves can be therefore described by observing the instantaneous displacement of the particles inside the medium. Spatial sound technologies, which is the focus of this dissertation, utilise multichannel signal processing tools for the capture, analysis, manipulation, and rendering of complex sound scenes.

2.1 Physical properties of a sound field

Obtaining information of sounds in a room can provide us with knowledge about the sound sources and the room itself. Sounds propagate in and interact with a medium, and the sound arriving at a receiving point comprises the direct sound, its early reflections and the late reverberation [3, 4]. In the context of this thesis, we consider the medium to be a closed or open space and that the basic properties of a sound field can be described by measuring the room impulse response (RIR) at a specific position inside a space. A RIR can be divided into three parts: the direct path, the early reflections that occur when the sound interacts with the boundaries of the medium, and the late reverberation. A typical RIR is shown in the top panel in Fig. 2.1. The direct path is the straight line connecting the sound source and the receiver, thus providing information

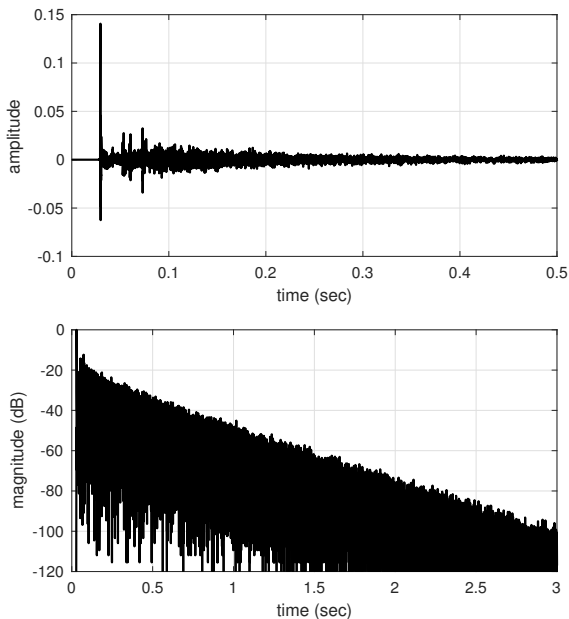


Figure 2.1. Impulse response of a concert hall (top) and the energy decay (bottom).

about the sound source itself. The early reflections are sound waves emitted from the source that interact with the walls, ceiling, floor or objects placed inside the room. The density of reflections increases over time, and after a certain amount of time, typically around 80 ms, they are considered to be part of the late reverberation (ISO 3382, 1997) [5]. The energy decay of the response, shown in the bottom panel in Fig. 2.1, is commonly characterised by the reverberation time RT_{60} which is defined as the time taken for the energy to decay 60 dB [3]. In the example in Fig. 2.1 (bottom) RT_{60} is approximately 1.5 sec.

A formula to estimate the reverberation time RT_{60} in seconds has been proposed by Sabine [3]

$$RT_{60} = 0.161 \frac{V}{A}, \quad (2.1)$$

where V is the volume of the room and $A = \alpha S$ is the equivalent absorption area with α being the average absorption coefficient and S the surface area. The assumption for estimating the reverberation time is that the decay of the energy starts from an ideal diffuse field, which can be approximated by the late reverberation [3].

Additional information about the room, such as the frequency content

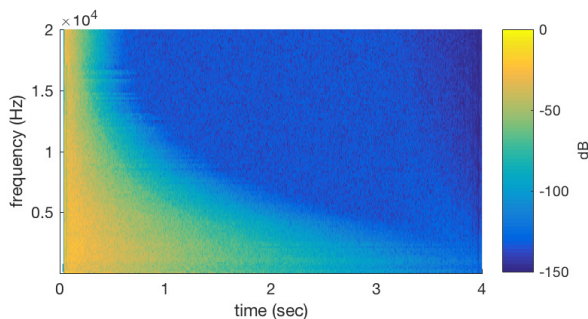


Figure 2.2. Frequency response over time of the RIR depicted in Fig. 2.1.

of the RIR in time, which is not evident in the RIR, can be obtained with a spectrogram. The room is excited by a sound source differently for different frequencies. Therefore a time-frequency analysis of the RIR can be informative. In Fig. 2.2, the spectrogram of the same RIR in Fig. 2.1 is depicted on a linear scale.

The physical properties of the sound field can be described by two quantities: the pressure p and the particle velocity \mathbf{u} . The pressure is scalar while the particle velocity is a vector. With these two quantities the instantaneous intensity vector with frequency k at a measurement point (x, y, z) is defined as

$$\mathbf{I}(k) = p(k)\mathbf{u}(k). \quad (2.2)$$

The real part of the intensity vector $\Re(\mathbf{I})$ (\Re is the real-part operator) is called the active intensity and indicates the direction of energy flow at the point [6, 7].

2.2 Capturing the properties of a sound field

The pressure variations caused by a sound source in a room can be captured at a point or at multiple points in space with acoustic transducers such as microphones. Microphones are used to measure the sound pressure and to transform it into an electrical signal. The two main types of microphones are the pressure and pressure gradient microphones. Pressure microphones comprise a membrane placed at the end of a cavity and, in principle, they can measure pressure variations equally from all directions [8]. In practice, and for high frequencies or small wavelengths, the

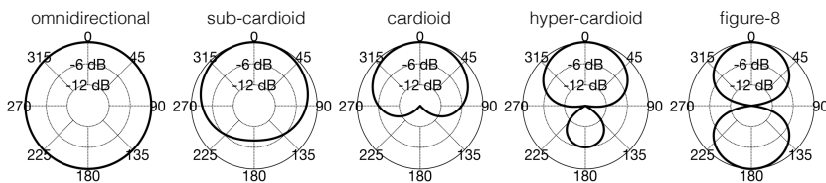


Figure 2.3. Polar graphs on a logarithmic scale of ideal first-order microphone.

cavity of the microphone affects the actual response of the microphone. This occurs when the microphone scaffolding or support is larger than the wavelength. Pressure gradient microphones, also known as a velocity microphones or figure-8 microphones consist of a diaphragm open from both sides. They are capable of detecting differences in pressure in both sides of the diaphragm. Pressure gradient microphones can also be constructed by subtracting the signals of two closely-placed omnidirectional microphones.

First-order directional microphones with varying directional patterns can be constructed by combining the signals of a pressure and a pressure gradient microphone with different weights and assuming that the microphones are placed at the same position [8]. The resulting directional pattern is described with the following polar equation

$$p_d(\phi) = w_p p(\phi) + (1 - w_p) u_x(\phi), \quad (2.3)$$

where p indicates the angular response of a pressure microphone for angle $\phi \in [0, 360]$, u_x is the angular response of a pressure gradient microphone, and w_p is a real-valued weighting factor. The resulting directivity patterns are shown in Fig. 2.3 for different values of w_p : sub-cardioid ($w_p = 0.7$), cardioid ($w_p = 0.5$) and hyper-cardioid ($w_p = 0.25$) [8].

2.3 Microphone arrays

Microphone arrays consists of a set of acoustic sensors positioned in different geometrical arrangements. The advantage of using multiple microphones is that the spatial attributes of the physical quantities of a sound field can be captured. Signal processing techniques that spatially sample the sound field have numerous applications, such as estimating the

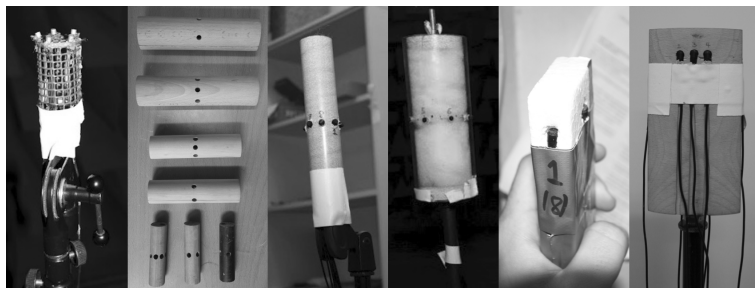


Figure 2.4. Prototype compact microphone arrays constructed for experiments during the dissertation. From left to right: open 5-channel circular microphone array or 1.5 cm radius, different-radii wooden rigid cylinders able to fit 5 to 8 microphones, 8-channel cylinder of 1.3 cm radius, 8-channel cylinder or 4 cm radius, 3-channel mobile-like array made with styrofoam and 7-channel mobile-like array made out of wood.

spatial attributes of a sound field [9, 10, 11, 12], localisation of single or multiple sources [13, 14, 15, 16], sound source separation or beamforming [17, 18] and sound field reproduction [19, 20, 21, 22, 23]. The underlying signal processing technique utilising microphone arrays is to combine the microphone signals so that at the output the signal is characterised by a directional selectivity. Such an operation is called beamforming. A number of signal-independent and signal-dependent techniques utilising microphone arrays are discussed in section 4.

There are two main categories of microphone arrays: additive arrays and differential arrays. Additive arrays were commonly utilised in signal enhancement and noise suppression applications and they usually require a large inter-microphone spacing. Each microphone signal measures and converts the acoustic pressure into an electrical signal. All the microphone signals are then combined in a signal processing unit. On the other hand, differential arrays measure the spatial derivatives of the acoustic pressure. This is performed by placing two sensors very close and subtracting the output signals to get a pressure gradient signal. By extending this concept to multiple microphones, higher order directivity patterns can be obtained but with the cost of microphone self-noise amplification [24].

Narrow directivity patterns can also be achieved by placing an omnidirectional sensor on the surface of a rigid object. Examples of such studies

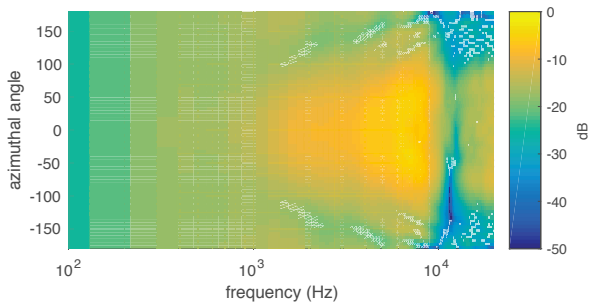


Figure 2.5. Angular frequency response of a single omnidirectional microphone fitted on the surface of a rigid mobile-like device.

are provided in [25]. Spherical and cylindrical baffles are of special interest since they provide a symmetrical directional response in two and three dimensions respectively. Analytical solutions for the directional patterns of microphones placed on rigid objects exist for spheres and infinitely long cylinders [26, 27]. For other arrangements one would need to either measure the steering vectors in an anechoic chamber or model them with numerical simulation software. By measuring the steering vectors in an anechoic chamber it is possible capture the exact geometry of the array and the characteristics of the different microphones.

A number of prototype compact microphone arrays have been developed within the scope of this work, shown in Fig. 2.4. An illustrative angular frequency response is depicted in Fig. 2.5. An angular spectrogram is essentially the frequency response of the microphone for a sound source in various directions. The response was measured in an anechoic chamber using the swept-sine technique [28]. The response is from a single omnidirectional microphone mounted in the middle of the the rightmost microphone array in Fig. 2.4. The effect of the rigid body is clearly seen in the figure as the response of the microphone becomes directional between 100 Hz and 10 kHz. Right after 10 kHz there is destructive interference which causes a dip in the response, and at all frequencies above that it becomes directional.

3. Perception

As discussed in section 2, sound is produced by the vibration of an object. These vibrations cause a disturbance in the surrounding particles. Specifically, the particles of the surrounding medium experience condensation and rarefaction, resulting in a pressure change. The particle movement is a local movement, producing a sound wave which moves outwards from the vibrating object and attenuates as it moves away. Sound is the human perception of vibrations in the region between 20 Hz and 20 kHz. In this section, the main aspects of spatial sound perception are discussed.

3.1 Human auditory system

The peripheral part of the human auditory system, which does not differ much in most mammals, is composed of the external ear, the middle ear, and the internal ear. An illustration is provided in Fig. 3.1. The external ear consists of three basic structures: the pinna, the ear canal and the eardrum. The main role of the pinna in hearing is the localisation and enhancement of certain frequencies. One important part of the pinna is the concha, responsible for boosting certain frequencies. Given that the shape of the pinna does not change dramatically, as does not its spatial orientation (in contrast to some other mammals), the efficiency with which it reflects sound towards the entrance of the ear canal depends only on the direction of arrival of a sound [1].

Sound travels inside the ear canal and causes the eardrum to vibrate. As the eardrum is near the inner end of the ear canal, it acts as a resonator. A tube closed at one end has a resonance frequency four times the length of the tube in meters. In the case of the ear canal this is approx-

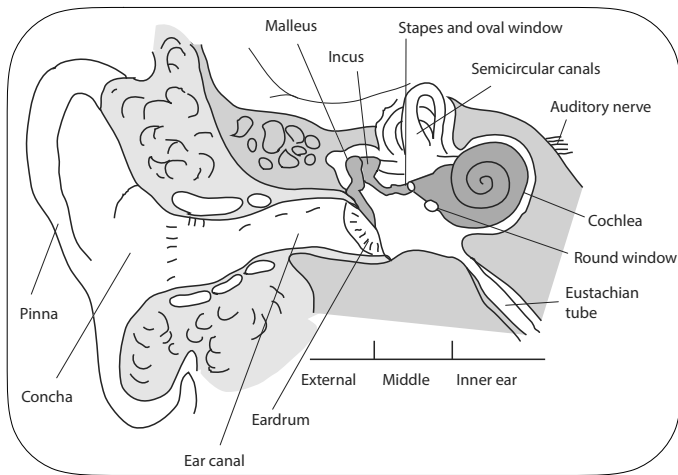


Figure 3.1. Human auditory system [1].

imately 3.5 kHz. Due to the fact that the eardrum vibrates and absorbs some energy, the resonance frequency of the ear canal is not exact and varies between 3-4 kHz. Due to the complexity of the structure of the eardrum, its vibration is quite complex, especially at high frequencies.

The eardrum separates the ear canal from the middle ear, which effectively connects the eardrum with the oval window through a set of bones, the ossicles. Muscles are attached to the ossicles (malleus and incus) and can attenuate intense sounds to the inner ear by construction, a phenomenon known as the acoustic reflex. This reflex reduces the amount of vibration transmitted to the oval window mostly in the low-frequency region. The oval window transforms the vibration into sound which is transmitted in the internal ear. The main function of the middle ear is to transmit the vibration motions of the ear drum to the sound receptors in the internal ear. In addition, it acts as a transformer. If the middle ear was absent, its most likely the incoming sound would be reflected back. Air is transmitted through the cavity of the middle ear through the Eustachian tube, which is normally closed but opens when swallowing or yawning. The role of the Eustachian tube is to balance the pressure in the middle ear with the pressure in the ear canal. An additional function of the middle ear is to reduce the transmission of internally generated bone-conducted sounds to the cochlea. Such sounds are produced by bone

vibration during chewing and can mask incoming sounds from the environment [29].

The inner ear consists of two parts, one responsible for the vestibular system, which contributes to balance and spatial orientation, and the other devoted to the auditory system, the cochlea. The cochlea is a hard bone with rigid walls, consisting of a spiral gradually decreasing in diameter. The entrance of the cochlea is through the stapes in the oval window. The cochlear spiral terminates at the apex and is divided longitudinally into three separate canals, the scala vestibuli, the scala tympani and the scala media, all filled with fluids. The cochlear spiral is divided by two membranes, the Reissner's membrane and the basilar membrane. On the basilar membrane resides the organ of Corti, an area containing hair cells that can be divided into two groups: the outer and inner hair cells. The inner hair cells act as transducers and transform the mechanical motion into neural activity. The role of the outer cells is to influence the mechanics of the cochlea by adjusting the sensitivity and the sharp tuning.

The information sent by the cochlea is received by the auditory nerve system with thousands of neurones transmitting it to the central nervous system. Studies of the auditory nerve system mostly consisted of placing micro-electrodes and measuring nerve impulses in auditory fibres [29]. Each fibre shows background activity whether or not there is an auditory event present. Additionally, the fibres show frequency selectivity, meaning that they are more sensitive to some frequencies. The frequency selectivity of a single fibre can be illustrated by a tuning curve. The characteristics of the tuning curves are usually defined by using tone bursts as stimuli and measuring the fibre's response. The frequency where the fibre has the lowest threshold is known as the characteristic frequency. The characteristics of a fibre above the threshold can be described by the isolated contours which can be measured and is the required intensity of a pure tone in order to produce a specific firing rate.

3.2 Time-frequency resolution of the human auditory system

Sound travels through the cochlea from the oval window to the apex and creates resonances at different places on the basilar membrane. When a

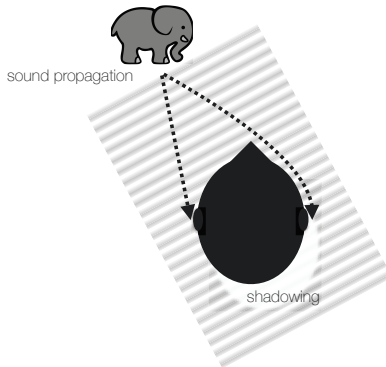


Figure 3.2. Sound arriving from a single source (norsu) to a listener. Human hearing exploits simultaneously different cues that contribute to determining the location of sounds. Sounds can arrive at the two ears with different delays and or different levels. The head shadowing and reflection from the pinna and torso provide spectra-temporal information.

wide-band signal reaches the eardrum, it sets in motion the bones in the inner ear which then set in motion the oval window. The pressure difference caused by this motion in the cochlea sets the basilar membrane in move. The response of the basilar membrane to sounds takes the form of a sound wave which moves the membrane and terminates at the apex. This type of wave increases gradually and decreases rapidly. Low-frequency sounds produce a vibration which reaches a maximum at the apex, but high frequencies have their maximum near the base of the membrane. The cochlea can be characterised as a Fourier analyser with a set of band-pass auditory filters with the centre frequencies of these filters positioned in different places on the basilar membrane. The shape and centre frequencies of these auditory filters have been studied in [30]. According to this study, the width of the auditory filters can be approximated with the equivalent rectangular bandwidth formula

$$f_{\text{ERB}} = 24.7(4.36f_c + 1), \quad (3.1)$$

where f_c defines the centre frequency of the filter in kHz.

3.3 Sound source localisation

3.3.1 Inter-aural time and level differences

The human auditory system employs a variety of sound localisation cues to determine the position of a sound source. An illustrative example is shown in Fig. 3.2. These cues are based on differences in the signals received by the two ears, which can be in time, intensity or spectrum. The most prominent cues for sound source localisation depend on the comparison of signals reaching the two ears. These are the inter-aural time difference (ITD) and inter-aural level difference (ILD). A comparison of the levels between the left and right ear refers to ILD. Humans have been shown to be extremely accurate in localising sounds emitting sinusoids of 500 Hz with accuracy as good as 1° [31, 1]. The ILD is a function of frequency over the whole audible range. Values of the ILD are relatively small for frequencies below 500 Hz since these frequencies correspond to wavelengths that are larger than the size of the head. For frequencies above 500 Hz, the wavelength is shorter and are diffracted by the presence of the head. When using pure tones, the cue of the ILD is most useful for high frequencies while the cue of the ITD is useful at low frequencies [29, 32].

3.3.2 Head-related transfer function

ITD and ILD alone are not adequate to describe localisation. For example, it is not possible to localise a noise source in the median plane solely with the ITD and the ILD. There are also spectral cues that affect localisation. These spectral cues are provided by the pinna, head and torso that affect the localisation of high-frequency sounds. The way the spectrum is modified can be described by the head-related transfer functions (HRTFs). Such cues are important in sound-source localisation for complex sound sources, as they can provide cues that resolve front-back confusion and/or determine elevation [32, 31].

3.3.3 Dynamic cues

Another important cue is head movements. If the head remains stationary, the given ITD and or ILD are not adequate for the unique localisation of a sound source placed along the cone of confusion [31]. The cone of confusion is a cone with its axis along the line formed with the two ears which corresponds to the same ITD and ILD. Head movements can reduce localisation ambiguity. Although there is large subjective variability concerning the effect of head movements, these movements have been shown to improve localisation and, in addition, to reduce front-back and vertical confusion [1].

3.3.4 Precedence effect

When a sound from a source reaches the human auditory system from one specific direction is followed rapidly by another sound originating from another direction, the perceived direction is dominated by the first sound source. This phenomenon is known as the law of the first wavefront or precedence effect. The perceived location of the fused sound sources depends on the size of the delay between the two arriving sound sources. Summing localisation occurs for delays between sound sources that are less than 1 ms while localisation dominance occurs when the location of the perceived fused sound source is determined by the first signal. The precedence effect is employed by the human auditory system to localise sounds despite the acoustical conditions. For example, it provides a cue to identify the location of a sound source in a reverberant environment [1].

3.3.5 Perception of distance

A number of factors determine the perception of distance, the ability of the human auditory system to perceive the distance of a sound source. For familiar sources it is possible to determine an approximate distance perception by judging the sound level. However, the perception of distance depends on whether the source and listener are positioned in open or enclosed spaces [1]. In open spaces, the sound level is the dominant cue. However, for long distances the sound level is not adequate, and the human auditory system employs changes in the spectrum as aids. These

changes correspond to the physical phenomenon of air absorption for high frequencies. However, judging the absolute distance is not possible, but only the relative distance of the sound sources, meaning which one is closer or further away. A different scenario occurs in enclosed spaces, where the direct sound is fused with diffuse sound due to the reflection paths between the source and receiver. Because of the precedence effect, the human auditory system fuses the reflected sound with the direct, but it is still possible to employ the information from the reflection to provide a sense of distance. Distance cues can be obtained from the direct-to-diffuse sound ratio and the spectrum of the reflected sound. It has been suggested that a weighting of these separate cues is performed depending on the source properties and environmental conditions [1].

3.4 Time-frequency resolution for perceptually-motivated signal processing

Knowledge of the capabilities and limitations of the human auditory system can provide information on how to apply signal processing techniques that are targeted for human listeners. Since humans analyse a sound scene in time and frequency, it seems logical that time-frequency domain signal processing techniques are applied. A common approach is to perform signal analysis with the short-time Fourier transform or filter-banks [33, 1]. For such time-frequency analyses the common requirements for perceptually transparent sound reproduction are that the time-frequency processing generates non-perceivable spectral aliasing between the adjacent frequency bands and that the resolution of the transform, whether it is common STFT or filter-bank-based, approximates the assumed resolution of the human auditory system [1]. The ERB bands provide a starting point. Several other filter banks have been proposed to approximate the resolution of the human auditory system in the literature, such as the ERBlet transform [34], the quadrature mirror filter-bank (QMF) used for spatial sound coding applications such as in [35] and in Publication IV, or the one implemented for the study in Publication II.

4. Directional filtering and analysis of spatial sound

Being able to focus on a sound source of interest in a noisy recording has triggered the interest of researchers in many different disciplines of the signal processing and acoustics community for several decades. Multiple microphones can be utilised in enhancement applications due to the increase in computational power. The type of signal enhancement that is of interest within the context of this thesis is directional filtering. In directional filtering, the task is to focus on a specific direction around the recording device. In this case, the assumption is that the noise is spatially distributed. Using multi-microphone techniques one could enhance the direction where the sounds of interest reside while suppressing other directions. The degree of enhancement and/or suppression depends on the number of microphones and the array configuration.

Multi-microphone devices enable flexible recording of sound sources in the presence of noise and reverberation. The most common enhancement techniques for microphone arrays are based on the design of directional filters. Directional filtering or beamforming with microphone arrays is a class of methods that allows focusing in specific directions in a sound scene. Such methods can be utilised for beamforming or, in case of multiple outputs that can cover the sound field, sound field reproduction, [17, 22, 36]. An illustration of beamforming applications is in Fig. 4.1. Multiple sound sources are simultaneously active in a room, and the task is to focus on one specific direction while suppressing interferers. Such algorithms are ideal candidates for general scenarios where high intelligibility and perceived quality is required, [37, 38, 39, 40]. Practical applications include teleconferencing [41, 42] or recording simultaneous instruments for music production [43].

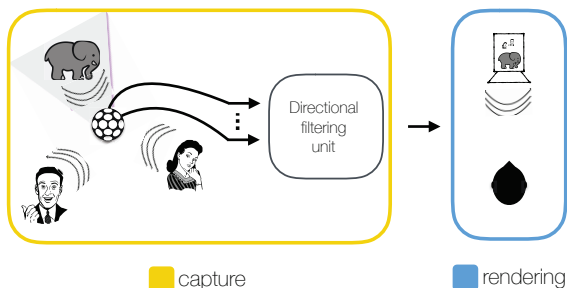


Figure 4.1. An illustrative scenario where multiple sources are captured with a microphone array. The task is to enhance sounds originating from the desired region (indicated with the darker triangle).

There is a plethora of techniques to approach the problem of directional filtering. The selection of the appropriate technique depends mainly on the application. For example, in cases where sound quality is the priority, the algorithm might be required to be less aggressive in terms of directional filtering. In cases where signal retrieval is the priority, audible artefacts that are generated by the algorithm might be of less importance. This section provides an overview of directional filtering techniques, such as beamforming and post filtering.

4.1 Signal-independent directional noise reduction

The most basic beamforming techniques are signal-independent and, as the name suggests, do not assume anything about the nature of the signals or the environment they reside in. A basic block diagram illustrating these simple operations is shown on the left in Fig. 4.2. The beamformer output y is calculated by applying a set of user-defined beamforming weights w_{ds} to the microphone signals $\mathbf{x} = [x_1, x_2, \dots, x_q]$ with an optional time-alignment block in Fig. 4.2, as in delay-and-sum (DaS) beamforming [44], or a set of weights w_{ls} estimated from the array-steering vectors and/or the array geometry, as in least-squares-based beamforming synthesis [45, 46, 47].

These techniques are based on performing simple operations between the microphone signals, such as adding or subtracting signals from spaced arrays [8, 38, 24]. The DaS beamformer algorithm estimates the time de-

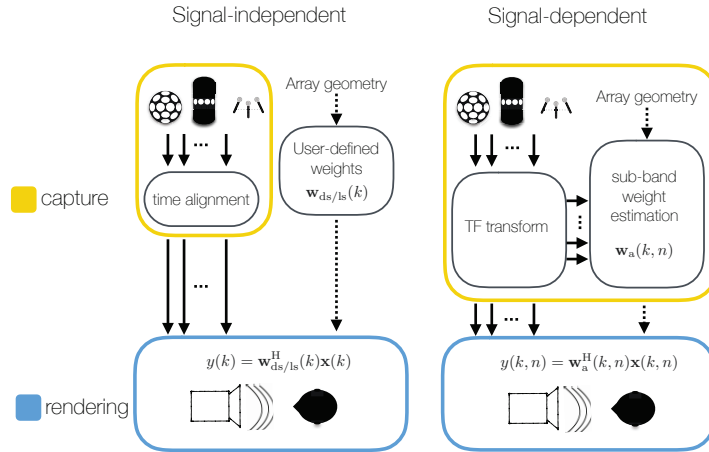


Figure 4.2. Beamforming: a signal-independent approach (left), where the signals are simply mixed with time-invariant weights $w_{ds/l_s}(k)$, and a signal-dependent approach (right) where sub-band weights $w_a(k, n)$ for time index n are applied to a time-frequency (TF) transformed signal from the microphones.

lays of signals received by each microphone of an array and compensates for the time difference of arrival [44]. By aligning and summing the microphone input signals, the directionality of the microphone array can be adjusted in order to create constructive interference for the desired propagating sound wave and destructive interference for sound waves originating from all other directions. Narrow directivity patterns can be obtained, but this usually requires a large spacing between the microphones. In a practical system, these constraints on the array size and the number of sensors results in a trade-off between directional selectivity, noise amplification and spatial aliasing.

A closely-spaced microphone array technique that uses least-squares minimisation has been proposed [48] and can be applied to beamforming for sound reproduction [22]. In this technique, the microphone signals are summed together with the same or opposite phase with different gains and frequency equalisation, where the target is a set of directional patterns following the spherical harmonics of different orders. Such techniques, also referred to as pattern matching, have been used in many applications where the target patterns can be HRTFs [49, 50], narrow

beamformers for sound recording [45, 51], or loudspeaker panning functions utilised in sound reproduction systems [51, 52, 36]. The resulting response has tolerable quality only within a limited frequency range. For compact microphone arrays there is a trade-off between directionality and noise amplification.

4.2 Signal-dependent directional noise reduction

4.2.1 Beamforming

Microphone arrays capture sounds inside an environment which are disturbed by noise and reverberation. Due to the linear superposition of acoustic waves, interference from other sources can be modelled as components added to the clean signal. Signal-dependent or adaptive beamformers, such as the linearly-constrained minimum variance (LCMV) or the minimum-variance distortionless response (MVDR) beamformers [17], operate in the time-frequency domain and steer a beamformer adaptively such that the noise is attenuated while the desired signal(s) are passed through with the user-defined gains. The basic block diagram for these techniques is illustrated on the right in Fig. 4.2. The techniques formulate complex-valued beamforming weights for each sub-band using constrained optimisation based on the short-time estimates of the microphone signal covariance matrix. The resulting beamformer weights minimise the energy of the output of the beamformer, thus suppressing the effects of the interfering sounds. Adaptive beamformer designs that constrain the microphone self-noise amplification may also result in reduced spatial resolution in the low frequencies, which is especially evident when utilising compact microphone arrays. As a consequence the ability to suppress the directionally distributed sound energy is degraded. This is especially true in typical acoustic conditions and at large wavelengths with respect to the array size.

Another class of adaptive beamformers are described as part of the informed spatial filters which combine beamforming with noise reduction. These adaptive filters use a model of the sound field based on the signal received by the microphone array. The sound-field model consists of a fi-

nite number of plane waves in addition to the diffuse sound component and the noise component. The effectiveness of this spatial filter relies on the accurate estimation of the noise statistics, the number and direction of plane waves, and the diffuse sound power. The studies indicate that such beamformers show good performance in acoustic conditions matching the sound field model [53, 54, 55, 56].

4.2.2 Post filtering

As discussed in the previous section, utilising adaptive beamformers might result in low directional selectivity which is evident in the directivity factor and is audible as a spectral imbalance and undesired reproduction of diffuse sound and noise in the output [57, 44]. Additional noise reduction, in a spectral sense, can be further achieved using post-filtering/masking techniques [58]. A block diagram is provided in Fig. 4.3 where the masker or post filter $G(k, n)$ is commonly applied to the output of an adaptive beamformer $\mathbf{w}_a^H(k, n)\mathbf{x}(k, n)$. Non-adaptive beamformers can also be utilised in applications that require low latency. The post-filter can be estimated from the input signals or predicted by a learning system that is trained using a set of features and target outputs [59, 60].

In traditional time-frequency masking approaches, the observed magnitude spectrogram is multiplied commonly with a real-valued post filter in order to remove noise and keep the desired signal. The key concept is to apply low values in the time-frequency regions that are dominated by noise and pass the target signal components unchanged. Masking is commonly applied at the output of the beamformer for adjusting the spectrum to match better that of the desired sound source [58]. However, artefacts from inaccurate post-filter estimation can result in short sparsely spaced peaks in the post-filter which can become perceptually evident when the output signal is transformed back to time domain as very short-time sinusoids that have been defined as musical noise [61]. Typical smoothing techniques, for example using a one-pole recursive filter design, and spectral floor adjustments have been applied to mitigate these artefacts.

Post-filter types based on noise estimates

One of the first post-filters in the literature for suppressing room reverberation was introduced in [62]. The design assumption of this post-filter

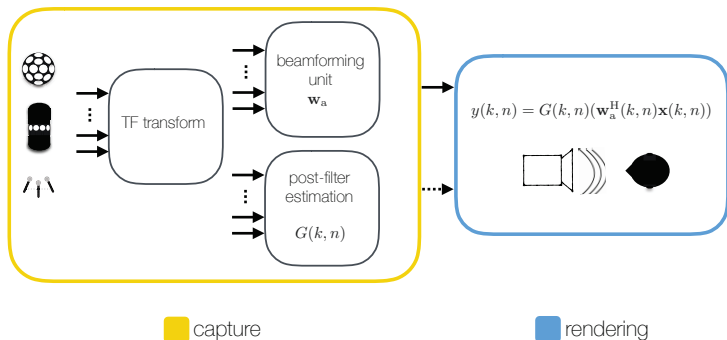


Figure 4.3. Post-filtering a beamformer

is that the noise in the sound scene received at each sensor of the microphone array is uncorrelated. Further work on this topic consists of an algorithm proposed in [63], where the sound field was modelled with a coherence function for a spherically isotropic field to identify correlated noise. The introduction of this post-filter provided a generalisation of the Zelinski post-filter [64]. These post filters are only capable of reducing uncorrelated or correlated noise in the beamforming output and rely on the output of the beamformer for the suppression of the directionally distributed interference. Unfortunately, these methods are characterised by poor performance at low frequencies when the correlation between microphone signals is high [44]. From a signal processing perspective, the optimal multi-channel filter in minimum mean squared error sense is the multi-channel Wiener filter, [48, 65] which has been shown to be equivalent to an MVDR beamformer followed by a single-channel Wiener filter [58].

Post-filtering based on spatial parameters

The assumption of the sparsity of the source signals is also utilised in another technique, directional audio coding (DirAC), which is a method to capture, transmit, and render spatial sound over arbitrary rendering setups. The most prominent DOA and the diffuseness of the sound field are measured as spatial parameters for each time-frequency position of sound. The DOA is estimated as the opposite direction of the intensity vector, and the diffuseness is estimated by comparing the magnitude of the intensity vector with the total energy. A variant of DirAC has been

used for beamforming [66], where each time-frequency position of sound is amplified or attenuated depending on the spatial parameters. In practice, if the DOA of a time-frequency position is far from the desired direction, it is attenuated in the rendering. However, in the cases where the assumption of sparsity is violated and two source signals are active at the same time-frequency position, intensity-based DOA provides inaccurate data, and artefacts may occur.

Recently, some techniques have been proposed which assume that signals arriving from different directions to the microphone array are sparse in the time-frequency domain, i.e. one of the sources is dominant at one time-frequency position [67]. Each time-frequency tile is then attenuated or amplified based on the spatial parameters analysed for the corresponding time-frequency position. A microphone array consisting of two cardioid capsules in opposite directions has been proposed in [68] for such a technique. Correlation measures between the cardioid capsules and Wiener filtering are used to reduce the level of coherent sound in one of the microphone signals. This produces a directive microphone, whose beam width can be controlled. An inherent result is that the width varies depending on the sound field. For example, with few speech sources in conditions of low level of reverberation, a prominent narrowing of the cardioid pattern is obtained. However, with many uncorrelated sources and in a diffuse field, the method does not change the directional pattern of the cardioid microphone at all. The method is still advantageous, as the number of microphones is low, and the setup does not require a large space.

A recent class of beamformers, which includes the technique described in Publication I, estimates the energy of the target signal using the cross-spectrum of two beam patterns with the constraints that their maximum sensitivity and equal phase are in the same look direction. This is illustrated in Fig. 4.4. When these constraints are met, the effect of the noise is suppressed whereas the energy of the source in the look direction is retained. The derivation of the post-filter in the cross-pattern coherence algorithm (CroPaC) relies on the calculation of the cross-spectrum of two static beam patterns. The application of the CroPaC is feasible with any order of microphone input, and the directional shape of the beam can be altered by changing the formation of the directional patterns of the micro-

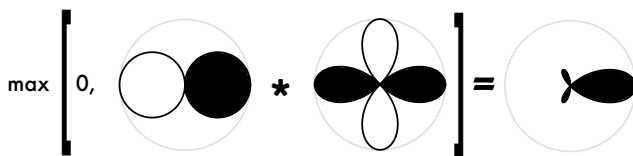


Figure 4.4. Illustration of the resulting directional selectivity of a post-filter derived by multiplying the output of two types of beamformers. Note that the beamformers have the same phase and equal magnitude in the same direction. The half-wave rectification process, shown as the maximum operation, ensures that any directional information arriving with negative phase is neglected. Adopted from [2].

phones from which the post-filter is computed.

A technique is proposed in Publication II to estimate a post-filter utilising the cross-spectrum technique of two beam patterns, where the selection of the beam patterns is based on minimisation techniques. The first beam pattern is static and corresponds to a spatially narrow beam pattern having equal spatial selectivity and unity gain across the frequency range of interest. The constant directional selectivity provides the grounds for spectrally balanced energy estimation in the presence of spatially spread noise. As a result of this design, this first beam pattern is characterised by high microphone self-noise gain at low frequencies, especially for compact microphone arrays [69]. The second pattern is formulated adaptively in time and frequency using constrained optimisation, with the constraint of suppressing the interferers while retaining the desired features of the orthogonality and unity zero-phase gain in the look direction. By these means, the proposed method inherits the noise-robust features of the prior technique in the class while providing the adaptive suppression of the discrete interferers. Using signals originating from a spatially selective but noisy beamformer for spatial parametric estimation has been previously shown to be effective also in the context of spatial sound reproduction for loudspeaker setups and headphones in [36] and [70], respectively. The derivation of the proposed algorithm is given in Publication II in the spherical harmonic domain, and the experiments are performed using a uniform spherical microphone array such that constant beamforming performance in all azimuthal and elevation positions is realisable.

4.3 Spherical harmonic analysis

Another type of beamforming that is relevant within the context of this thesis is based on a spherical harmonic analysis of a sound field. The background of this process is based on solving the wave equation in spherical coordinates [71, 27]. From a signal processing perspective, it is essentially a beamforming operation where a set of complex weights is applied to the microphone signals, typically in a spherical arrangement, in order to obtain a set of coincident beamformers that follow the directional selectivity of a set of orthogonal basis functions, the spherical harmonics [27]. The motivation to use SMAs for sound-field capture and rendering is that they are characterised with a similar performance in all directions when sensors are placed uniformly or nearly-uniformly on a sphere.

Spherical harmonic signals are essentially an intermediate representation between the microphone signals and the potential application. They also provide a convenient format for sound-field manipulation. For a detailed overview of these methods, the reader is referred to [21, 22, 27, 72, 26]. Let us denote a microphone array with Q microphones. A common approach is to decompose the microphone input signals $\mathbf{x} \in \mathbb{C}^{Q \times 1}$ into a set of spherical harmonic signals \mathbf{s} for each frequency. The accuracy of this decomposition depends on the microphone arrangement, the type of the geometry of the array [27]. The total number of microphones and their arrangement define the highest order of spherical harmonic signals L that can be estimated.

The sub-band spherical harmonic signals can be estimated as

$$\mathbf{s}(k) = \mathbf{W}(k)\mathbf{x}(k), \quad (4.1)$$

where

$$\mathbf{s} = [s_{00}, s_{1-1}, s_{10}, \dots, s_{LL-1}, s_{LL}]^T \in \mathbb{C}^{(L+1)^2 \times 1} \quad (4.2)$$

are the spherical harmonic signals and $\mathbf{W} \in \mathbb{C}^{(L+1)^2 \times Q}$ is the frequency-dependent spatial encoding matrix. For uniform and nearly uniform microphone arrangements, the encoding matrix can be calculated as

$$\mathbf{W} = \alpha_q \mathbf{W}_l \mathbf{Y}^\dagger, \quad (4.3)$$

where α_q are the sampling weights, which depend on the microphone distribution on the sphere [27]. $\mathbf{W}_l \in \mathbb{C}^{(L+1)^2 \times (L+1)^2}$ is an equalisation matrix

that eliminates the effect of the array geometry. A summary of the different schemes to calculate the equalisation matrix \mathbf{W}_l is shown in [73, 2]. $\mathbf{Y}(\Omega_q) \in \mathbb{R}^{Q \times (L+1)^2}$ is a matrix containing the spherical harmonics

$$\mathbf{Y}(\Omega_q) = \begin{bmatrix} Y_{00}(\Omega_1) & Y_{00}(\Omega_2) & \dots & Y_{00}(\Omega_Q) \\ Y_{-11}(\Omega_1) & Y_{-11}(\Omega_2) & \dots & Y_{-11}(\Omega_Q) \\ Y_{10}(\Omega_1) & Y_{10}(\Omega_2) & \dots & Y_{10}(\Omega_Q) \\ Y_{11}(\Omega_1) & Y_{11}(\Omega_2) & \dots & Y_{11}(\Omega_Q) \\ \vdots & \vdots & \vdots & \vdots \\ Y_{LL}(\Omega_1) & Y_{LL}(\Omega_2) & \dots & Y_{LL}(\Omega_Q) \end{bmatrix}^T, \quad (4.4)$$

where Y_{LL} are the spherical harmonic functions [71, 27].

Beamforming in the spherical harmonic domain can be expressed as

$$y(\Omega_o) = \mathbf{w}_{\text{PWD}}^H \mathbf{s}, \quad (4.5)$$

where Ω_o is the look direction and $\mathbf{w}_{\text{PWD}} \in \mathbb{C}^{(L+1)^2 \times 1}$ is a vector containing the steering vectors

$$\mathbf{w}_{\text{PWD}} = \mathbf{y}(\Omega_o) \odot \mathbf{d}, \quad (4.6)$$

where $\mathbf{y}(\Omega_o) \in \mathbb{C}^{1 \times (L+1)^2}$ is a row of the spherical harmonics matrix as denoted in (4.4), \odot denotes the Hadamard product and \mathbf{d} is a vector of weights defined as

$$\mathbf{d} = [d_0, d_1, d_1, d_1, \dots, d_L] \in \mathbb{R}^{1 \times (L+1)^2}. \quad (4.7)$$

The weights \mathbf{d} can be adjusted to synthesise different types of axis symmetric beamformers: regular [27], in-phase [21], maximum energy [74, 22] and Dolph-Chebyshev [27]. A comparison of the performance of such beamformer as DOA estimators is presented in Publication V.

Let us denote the covariance matrix of the spherical harmonic signals as $\mathbf{C}_{\text{lm}} \in \mathbb{C}^{(L+1)^2 \times (L+1)^2}$. The covariance matrix can be estimated using an average over finite time frames, typically in the range of tens of milliseconds or by recursive schemes. Signal-dependent beamforming in the spherical harmonic domain can be, for example, formulated as a MVDR minimisation problem. The synthesised beamformer adaptively changes according to the input signal, and its response is constrained to unity in the look direction while minimising the variance of the output [27]. By omitting the time and frequency indices, the minimisation problem is defined as

$$\begin{aligned} & \text{minimise } \mathbf{w}_a^H \mathbf{C}_{\text{lm}} \mathbf{w}_a \\ & \text{subject to } \mathbf{w}_a^H \mathbf{y} = 1. \end{aligned} \quad (4.8)$$

The resulting weights are

$$\mathbf{w}_a = \frac{\mathbf{y}^H \mathbf{C}_{lm}}{\mathbf{y}^H \mathbf{C}_{lm} \mathbf{y}}. \quad (4.9)$$

An advantage of using the MVDR in the spherical harmonic domain instead of the space domain is that the steering vectors are simply the spherical harmonics for different angles. A power map can then be calculated by using the output of the beamformer for different directions around the microphone array.

4.4 Direction-of-arrival estimation

Beamforming is a fundamental block in DOA estimation. DOA estimation is a well-studied class of parameter estimation algorithms with a wide selection of algorithms, such as subspace [14, 75, 76], intensity-based [77, 78, 79, 80, 81, 82, 83, 84, 85], and power spectrum methods [86, 87], each of them with a different level of complexity. The choice of the algorithm depends on the requirements of the application: the tolerable latency and the required accuracy. Steered-response beamforming and subspace methods such as MUSIC can provide accurate DOA estimates and have been extended in three dimensions for spherical microphone arrays [88]. However, the estimate requires an exhaustive search and therefore can be computationally heavy for real-time applications.

Active intensity-based methods utilise a pressure and a particle velocity component to analyse the sound field. In practice the pressure and 3-D particle velocity are estimated with a single omnidirectional and three dipole microphones, respectively [89]. Due to its tolerable latency, the intensity vector is an ideal candidate for low-latency DOA estimation and has been previously employed in time-frequency domain spatial sound processing [23]. Its performance has been examined in reverberant environments [85], and the formulation has been extended in the spherical harmonic domain in the form of the pseudo-intensity vector [79, 81]. The pseudo-intensity vector has been studied and compared with steered-response power beamformers and is an effective alternative in DOA estimation due its low computational complexity.

In Fig. 4.5, an example block diagram is shown where the active intensity is utilised as an instantaneous time-frequency DOA estimator. The

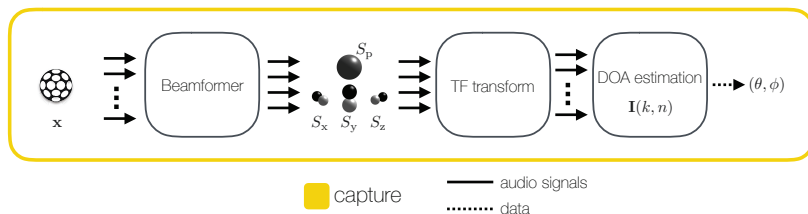


Figure 4.5. Block diagram for DOA estimation using the active intensity.

microphone array signals are fed into a beamforming unit where an omnidirectional and three dipole beamformers are synthesised. The omnidirectional beamformer approximates the pressure and the three dipoles the particle velocity. The active intensity is estimated for each time-frequency tile as

$$\mathbf{I}(k, n) = -\frac{1}{2} \Re \left\{ S_p(k, n)^* \begin{bmatrix} S_x(k, n) \\ S_y(k, n) \\ S_z(k, n) \end{bmatrix} \right\}, \quad (4.10)$$

where S_p , S_x , S_y and S_z approximate the pressure and particle velocity for the x- and y- and z-axis, respectively, of the Cartesian coordinate system. Since the active intensity indicates the energy flow for a given time-frequency tile, the DOA can be estimated as the direction opposite to that of the active intensity vector [90].

A higher-order active intensity has been proposed for parametric sound-field reproduction where the pressure and particle velocity are estimated in spatially constrained regions [91]. This is essentially performed by applying a spatial window to the omnidirectional and dipole beamformers. This concept has been exploited as a DOA estimator for an array fitted on a mobile-like device, and its formulation and performance have been examined in Publication VI.

4.5 Acoustic camera for sound field visualisation

Beamformers can also be utilised as part of an acoustic camera. In principle, an acoustic camera provides an acoustical analysis of a sound field. The result can be seen as a power-map overlaid on top of an image indicating the position and relative level of sounds. Power maps are useful

for applications such as sound field analysis, visualisation or simply for tracking sound sources. Incorporating this visual information in a power-map can significantly improve the understanding of the surrounding environment, such as the location and spatial spread of sound sources and potential reflections arriving from different surfaces.

Current techniques that utilise conventional beamformers perform poorly in reverberant and noisy conditions, due to wide width of the main lobe but also the side-lobes of the beamformers used for the power-map. A real-time acoustic camera is implemented as part of this thesis utilising signals-independent and signal-dependent beamformers but also using MUSIC and CroPaC algorithms. In Publication VII a probability-like parameter of sounds appearing at specific locations utilising the principles of CroPaC is developed. A side-lobe suppression algorithm is proposed in the Publication VII based on the product of multiple CroPaC parameters. The capture and visualisation is performed utilising a spherical microphone array and subsequently estimating a parameter to determine sound source activity at specific directions. A spherical camera is utilised to capture the visual field. In addition to the acoustic camera software a real-time software for transforming the microphone array signal into spherical harmonic signals is available here ¹.

¹<http://research.spa.aalto.fi/publications/papers/acousticCamera/>

5. Reproduction of spatial sound

5.1 Overview

The objective of a spatial sound reproduction system is to capture, transmit, and render a sound scene as close by as possible to the original in case of a real recording or the intended one in case of a synthesised/virtual sound environment. It is a single or multiple-input and multiple-output system, where the input consists of a monophonic or microphone array signal(s) and the output of a loudspeaker array or headphone signals. Figure 5.1 depicts the concept of such a system. Sound reproduction can be seen from two different perspectives. The first of them is signal-independent or non-parametric, where the capturing, transmitting and rendering is the same for any kind of input signal. The second is signal-dependent systems where the capturing, transmission, and rendering depend greatly on the sound scene: the position of the sound sources and how the sounds evolve over time, their frequency content, and the space they reside in.

5.2 Signal-independent reproduction

The main assumption in signal-independent techniques is that a sound scene can be captured with a set of beamformers with ideal capture patterns that depend on the reproduction setup. In the case of loudspeakers, such capture patterns can be frequency-independent panning functions, and in the case of headphones, a set of HRTFs. These beamformers are derived from a microphone array. A common downside is the low spatial

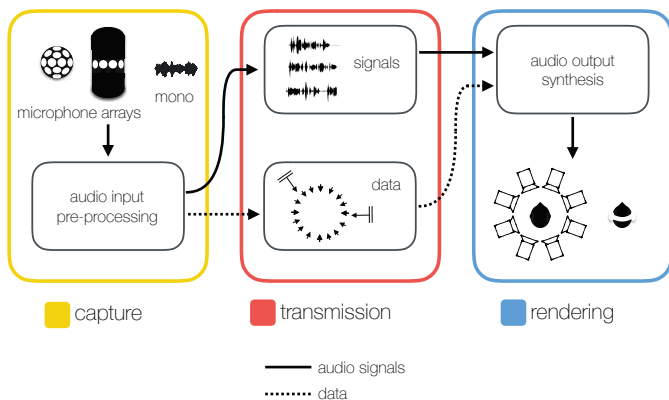


Figure 5.1. Generalised components of spatial sound reproduction: capture, transmission and rendering.

resolution, which is especially evident when capturing a sound scene with compact microphone arrays. The result is a blurred sound image and frequency colouration that can affect the perceived sound quality [92, 93, 94].

A signal-independent system can reside in one the following categories or any combination of them: channel-based, object- or panning-based, and scene-based. Each one of them offers different features and advantages and depends on the requirements of the immersive audio application and intention of the researcher or producer. These are depicted in Fig. 5.2 and analysed in detail below.

5.2.1 Channel-based systems

The class of channel-based spatial sound reproduction techniques is the most straightforward. It utilises the microphone-array input signals and feeds them directly into the output setup (see Fig. 5.2, left). A generalised microphone-to-loudspeaker relation is

$$\mathbf{y} = \mathbf{w}_{\text{eq}} \odot \mathbf{x}, \quad (5.1)$$

where $\mathbf{y} \in \mathbb{R}^{Q \times 1}$ are the loudspeaker or headphone signals, $\mathbf{x} \in \mathbb{R}^{Q \times 1}$ the microphone input signals, and $\mathbf{w}_{\text{eq}} \in \mathbb{R}^{Q \times 1}$ a set of optional equalisation filters that can be applied to each microphone signal separately. The principles behind such techniques for two or more channels are explained in [8, 95, 96]. For stereophonic reproduction, traditional two-channel mi-

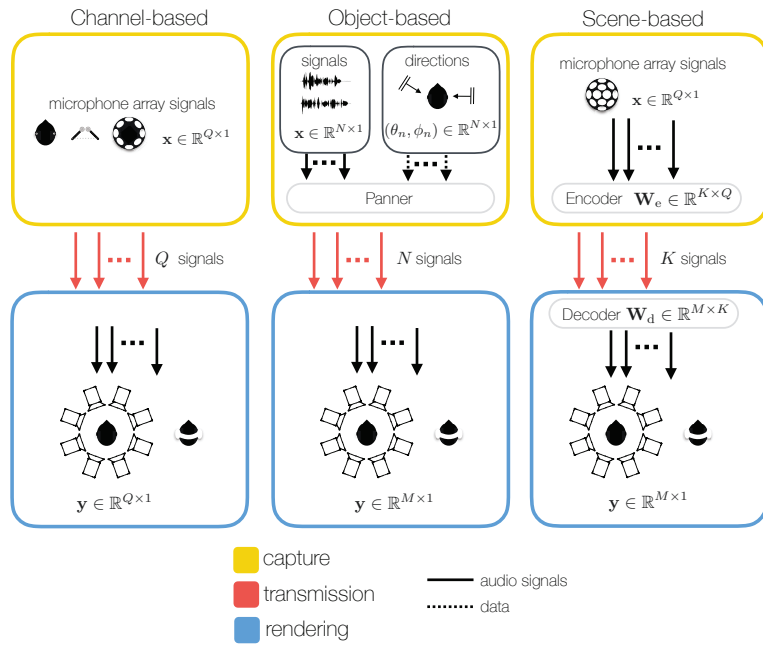


Figure 5.2. State-of-the-art signal-independent spatial sound reproduction systems: channel-based (left), object-based (middle) and scene-based (right).

crophone recording techniques were proposed, such as the A/B , X/Y , office de radiodiffusion télévision Française (ORTF), the Blumlein pair, and the mid-side stereo recording [8]. Each technique requires a specific microphone directivity ranging from omnidirectional to figure-8 and specific microphone and loudspeaker placement. Any deviation in the input and output setup degrades the rendered quality. In addition, due to the frequency-dependent microphone directivity, there is usually high correlation between signals at low frequencies, which blurs the perceived audio image. The popularity of these techniques has reached surround setups such as the conventional 5.1 or 7.1, with dedicated microphone arrays [97]. In binaural reproduction, channel-based techniques either utilise a simplified spherical head with two microphones or a dummy head for recording and feeding the output channels directly to headphones. Additional issues with personalised headphone-based audio, due to the individualisation of HRTF, are caused by such techniques. They can be partially solved by using binaural microphones placed at the ears of a human subject but with limited scalability. Although there are many downsides in channel-based techniques the main advantage of these techniques, causing many audio engineers to use them, is that no complex encoding or decoding process are involved, and the production can be performed by directly mixing the microphone signals.

5.2.2 Object or panning-based systems

A technique that can adapt to different loudspeaker/headphone setups is the object-based reproduction technique. Objects are considered monophonic signals accompanied by side data (see Fig. 5.2, middle). The objects are transmitted, and they are mixed at the rendering stage. The side data is usually the position of each object over time, which can be described with panning gains. Many panning techniques are available for such tasks, for example [98]. No prior information is required in the capturing stage, and the output setup can be taken into consideration only at the rendering stage. The input-output relation in this case is

$$\mathbf{y} = \sum_{n=1}^N \mathbf{W}_{\text{pan}}(\theta_n, \phi_n) \mathbf{x}_n, \quad (5.2)$$

where $\mathbf{W}_{\text{pan}}(\theta_n, \phi_n)$ are the panning gains for each sound source \mathbf{x}_n , and N is the total number of sources.

5.2.3 Scene-based systems

More versatile systems, such as the scene-based depicted on the right in Fig. 5.2. decompose the microphone input signal into an intermediate format which can be then decoded, in principle, to any output setup as long as the information about the position of the loudspeakers is provided. The microphone input-to-output relation can be described as

$$\mathbf{y} = \mathbf{D}\mathbf{R}\mathbf{E}\mathbf{x}, \quad (5.3)$$

where \mathbf{E} is an encoding matrix, \mathbf{R} is a sound scene manipulation matrix (for example, a rotation matrix), and \mathbf{D} is the decoding matrix. The flexibility of this system is that once the signals are encoded ($\mathbf{E}\mathbf{x}$) they can be transmitted, and the decoding is then independent of the microphone input signals.

First-order ambisonics (FOA) and higher-order ambisonics (HOA) are such systems [20, 21, 99, 22, 52]. The acoustical background of this technique is based on solving the wave equation in the spherical coordinate system [71, 27]. From a signal processing perspective, the microphone signals are decomposed into a set of spherical or cylindrical harmonic signals which are essentially a set of coincident beamformers following the directivity of various sets of basis functions that depend on the array geometry and specification [71]. Less flexible beamforming techniques have been proposed that synthesise frequency-independent beamformers in the loudspeaker directions or approximate the frequency varying HRTFs for headphone reproduction [100, 101, 102, 103]. They require, however, information about the input and output setup. In this case, the three matrices in (5.3) are combined into a single transformation matrix.

Although such systems can be used with different reproduction setups, they commonly suffer from high inter-channel coherence between the output signals, especially for low-order systems and for practical, compact microphone arrays. This leads to a perceivable sound colouration and low-frequency amplification of the original sound scene. However, they are typically characterised by high single-channel quality.

5.3 Parametric reproduction

Parametric techniques for spatial sound reproduction are traditionally based on a description of the sound field with a set of parameters which are measured adaptively in time and frequency. In contrast to signal-independent approaches, the parametric approach depends on the type of sound sources, the position and movement of the sound sources within the sound scene, and the space itself. A set of parameters is estimated for each time-frequency bin and is then used to render the input signals to loudspeakers or headphones. These parameters can be, for example, DOA or a direct/ambience decomposition of the sound field. Such systems require a small number of microphones and are very efficient in reproducing arbitrary sound scenes. Deviations from the parametric model might lead to parameter estimation errors, which in turn leads to reproduction artefacts. Artefacts are time-variant and usually more annoying than the frequency colouration of signal-independent systems. Such artefacts of the system, which are often described as bubbling or wobbling, are caused by amplitude and or direction modulation in frequency bands due to inaccurate parameter estimation. However, greater spatial resolution and better perceived quality can be achieved using state-of-the-art parametric methods compared to non-parametric techniques [2, 104, 105, 33, 106].

Parametric spatial audio processing techniques were initially introduced as data-compression algorithms, similar to the concept of MPEG-1 Layer 3 (MP3) and advanced audio coding (AAC), [107] but for coding multi-channel audio signals. A plethora of articles and doctoral theses have been published in the recent years on parametric spatial audio processing, for example [2, 105, 33, 90, 108]. The most obvious motivation behind the parametric, non-linear or time-frequency domain spatial-sound reproduction systems is data compression since multichannel recordings from a microphone array can be down-mixed and transmitted along with a set of spatial parameters. Exploiting the limitation of human hearing and analysing only the properties of the sound field that are relevant to human listeners provides more meaningful signal processing techniques. The rendering then delivers the necessary spatial cues to provide an engaging aural experience, perceived as close as possible to the recorded sound scene rather than a perfect reconstruction of the sound field.

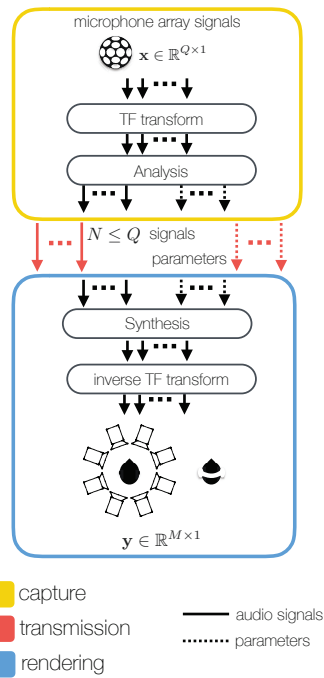


Figure 5.3. Parametric spatial-sound reproduction. Time-frequency representations of microphone array signals are analysed in the capturing stage, the microphone signals or a down-mixed signal of them is transmitted along with a set of parameters. At the rendering stage the captured sound scene is synthesised for an arbitrary loudspeaker or headphone setup.

5.3.1 Inter-channel level difference, time difference and coherence

One of the first approaches in parametric spatial-sound reproduction was BCC to compress binaural signals [109, 110, 111]. The input to BCC is at least two-channel. The audio input is transformed into the time-frequency domain. The main assumption is that the output signal should approximate the original or intended binaural spatial cues. The binaural cues considered within the context of BCC are ILD, ITD and later ICC [110]. The estimated parameters are the inter-channel level difference (ICLD), inter-channel time difference (ICTD), and inter-channel correlation (ICC) between the output signals. For headphone reproduction these parameters are identical to the binaural cues. The assumption is that these parameters preserve the binaural cues in the rendering stage. The parameters estimated for each frequency k are

$$\text{ICLD}(k) = 10 \log_{10} \frac{P_{x_N}}{P_{x_1}}, \quad (5.4)$$

where P_{x_N} and P_{x_1} are short-time estimates of the power of the signals x_N and x_1 , respectively,

$$\text{ICTD}(k, n) = \arg \max_d [\Phi_{1N}(d, k)], \quad (5.5)$$

where $\Phi_{1N}(d, k)$ is a short-time estimate of the cross-correlation function

$$\Phi_{1N}(d, k) = \frac{P_{x_N x_1}(d, k)}{\sqrt{P_{x_N}(k - d_1) P_{x_1}(k - d_2)}}, \quad (5.6)$$

where $P_{x_N x_1}$ is a short-time estimated of the mean of $x_N(k - d_1)x_1(k - d_2)$ and

$$d_1 = \max -d, 0, \quad (5.7)$$

$$d_2 = \max d, 0. \quad (5.8)$$

These parameters are then transmitted along with monophonic down-mixed input signals. In the synthesis stage the monophonic signal is duplicated for the number of output channels and manipulated with the parameters. The ICLDs are synthesised by multiplying the signals with factors that have ratios equal to ICLDs. The ICTDs are synthesised by applying delays equal to ICTDs and the ICCs by adding randomisation to the ICLD values [109]. ICLDs, ICTDs and ICCs can also be estimated from the covariance matrix of the input channels, as for example,

is shown in [33, 36]. The diagonal terms contain all the channel energies and the off-diagonal terms the inter-channel dependencies. Extraction of time or phase, level differences, and coherences between the channels are assumed to capture the statistical signal dependencies that need to be preserved and recreated. A technique similar to BCC is the parametric stereo (PS) which is always two-channel [112]. The main difference is in the parameter estimation, where PS utilises the inter-aural phase difference (ICPD) instead of the ICTD.

5.3.2 Direction of arrival and sparsity assumption

A simplified model of the sound field, where a single DOA per time frequency bin is utilised as a parameter, is proposed for binaural reproduction in [113]. In this single-plane-wave assumption the DOA is estimated from the phase differences of a compact four-microphone array. The DOA and a single microphone signal are then sent to a synthesis engine to render them using HRTF data. The main assumption here is that speech and music signals are usually sparse in the time-frequency domain [114]. The disjointness of the sound sources can be estimated with the W-Disjoint Orthogonality (WDO). Speech signals are usually mixed randomly whereas music-signal mixes depends on the type of music. Tonal music has a higher probability of overlap in the time-frequency domain. In this case the disjointness depends highly on the time-frequency analysis parameters.

An extension to the single-plane-wave model is utilised in HARPEX, where two active plane waves are assumed per time-frequency bin [115]. The DOAs are estimated by using matrix decomposition of signals from a tetrahedral microphone array. A multiple-DOA and diffuse model has been proposed in [116] using the same array. Adaptive beamforming techniques for various microphone array geometries has been proposed in [117, 118, 100].

5.3.3 Direction of arrival and diffuseness

Based on principles similar to BCC or PS, DirAC is a perceptually motivated parametric reproduction method that assumes that for a convincing reproduction for human listeners there is no need for a physically accu-

rate model [23]; Preserving the perceptually significant properties of the original is sufficient. The main idea was initially introduced to process impulse response and was named spatial impulse response rendering (SIRR) [89, 119]. The main assumptions used in DirAC are summarised below.

- A single DOA per time and frequency bin provides the ITD, ILD and monaural cues in the reproduction. These cues affect timbre.
- The diffuseness provides the ICC cues in the reproduction.
- The perceived auditory spatial image is determined by the DOA, the diffuseness, and the spectrum of the measured sound field.

The estimated parameters are the DOA and diffuseness, which are usually measured with a tetrahedral microphone array. The DOA is estimated using the active intensity, as discussed in section 4.4. The pressure and particle velocity are estimated with an omnidirectional and three orthogonal dipole microphones, respectively. The diffuseness is estimated from the following ratio [89]

$$\psi = 1 - \frac{\|\mathbf{E}[\mathbf{I}]\|}{cE[E]}, \quad (5.9)$$

where \mathbf{I} is the active intensity vector and E the energy density.

There are many variants of DirAC using different microphone arrays, such as tetrahedral, planar, cylindrical, spaced [120, 121, 122, 123]; using different analysis methods [124, 125, 126, 127]; using different synthesis methods [128]; using optimal mixing methods [129] and using higher order active intensity vectors [91]. It has also been a subject of many theses [90, 104, 105, 33, 130].

5.3.4 Target-setup parametrisation with optimal mixing

A beamformer-based parametric method is discussed in this work, which, instead of estimating the intermediate sound-field related parameters, estimates directly the perceptually relevant parameters of the reproduction setup. The rendering setup can be either a loudspeaker array or headphones. A set of target beamformers for the loudspeaker setup are defined as panning functions. For the case of headphone reproduction a set

of HRTFs is used. The system is based on two sets of beamformers for the parametric analysis and synthesis of a sound field. A set of narrow and potentially noisy beamformers is used in the analysis stage to estimate the target setup parameters, such as inter-channel coherences and energies, and a second set of beamformers, which are noise robust are used as the source signal. The method utilises optimal mixing techniques to enhance the robust beamformers with the directional characteristics of the narrow beamformers so that the relevant perceptual cues are preserved. The theoretical background and evaluation is described in detailed in Publications III and IV. The advantages of the proposed method are the reproduction of multiple instantaneous sound sources, the improvement of the single-channel audio quality, and the use of compact microphone arrays.

5.3.5 Object-based parametrisation

Signal-independent panning techniques have been extended in the parametric domain, where all objects can be encoded, down-mixed, transmitted and rendered. The separate sound sources can be, for example, extracted from a multichannel recording along with the side metadata. Such a framework has been proposed by the motion pictures experts group (MPEG) audio group [131]. For dynamic environments, this class of spatial sound reproduction can provide the flexibility of adjusting the position of sound objects with user-defined controls, which is especially attractive in games and in virtual reality applications [132, 133].

Primary ambient extraction techniques can be used for the task on decomposing a sound scene into separate objects. Principal component analysis is a popular technique for this task, whose performance depends on the correlation between the different components [134]. The decomposition task can be also performed by utilising blind source separation (BSS) separation techniques, assuming that the objects in the multichannel recording are mutually uncorrelated or statistically uncorrelated [135].

6. Summary of contributions

6.1 Directional filtering utilising the cross-pattern coherence

Publications I considers the task of directional filtering and presents a novel idea on how to use coherence-inspired measures between different types of beamformers to focus on sound sources and decrease the level of interferers and diffuse noise. The beamformers are signal independent, and they are estimated by utilising either measured or simulated microphone array-steering vectors. A parameter is estimated based on the cross-spectrum between two beamformers. The parameter, essentially a time-frequency soft masker, is normalised by the energy of the beamformers to establish values between zero and one, thus ensuring a distortionless output. The soft masker is then applied to the output of a beamformer with omnidirectional characteristics. The algorithm was evaluated in both simulated and real environments. A prototype cylindrical array was build to evaluate the performance of the algorithm in a reverberant environment. A second parameter related to the perceived quality of the algorithm is the spectral floor which sets the lowest value of the post-filter. As a general rule, low spectral-floor values provide the highest amount of attenuation but at the same time they might cause audible artefacts, such as musical noise. A perceptual evaluation was conducted to define the spectral-floor value for which the quality of the output is closest to a reference signal. The performance of the algorithm was shown to improve noise reduction when compared to the previous state-of-the-art.

6.2 Directional filtering based on orthogonally-weighted beamformers in the spherical harmonic domain

Publication II builds on the idea of utilising the cross-spectrum between two beamformers using the concept of orthogonally-weighted beamformers for general microphones array signals. The algorithm uses cross-spectral estimates between a static and an adaptive beamformer to formulate a time-frequency soft masker. The weights for the static beamformer are designed so that the output is characterised by constant directivity. This potentially causes noise amplification for compact microphone arrays. However, noise amplification is not audible in the output of the system since the beamformer signals are only utilised for parameter estimation and not for generating audio. The weights for the adaptive beamformer are estimated with an orthogonality constraint with respect to the weights of the static beamformer and unity gain in the look direction. These constraints provide distortionless response, diffuse noise and interferer suppression. The soft masker is then estimated by estimating the cross-spectrum between the output signals of these two beamformers. The cross-spectrum provides the target energy in a given look direction. The algorithm was evaluated with instrumental measures and with listening tests, and its performance was found to be superior when compared to the previous state-of-the-art.

6.3 Perceptually-motivated spatial sound reproduction based on optimal mixing

Publication III proposes a novel idea for perceptually-motivated sound reproduction. The perceptual aspects of the sound field are captured with a set of beamformers, the analysis beamformers, that match the ideal panning functions with respect to the target loudspeaker setup. This is a common cause of noise amplification in signal-independent systems and for compact microphone arrays. However, these signals are only utilised for parameter estimation. A second set of beamformers, the synthesis beamformers, are also estimated that match the target loudspeaker panning function but in a noise-robust manner. This results in beamformers being less directionally selective at low frequencies and potentially

cause high inter-channel coherence if reproduced directly from the output setup. An optimal adaptive mixing process is then utilised to exploit the different advantages of these two separate types of beamformers. The synthesis beamformers are enhanced with a parametric analysis of the directionally selective beamformers. One of the main underlying assumptions of this technique is that the necessary spatial cues, such as the inter-channel time difference, the inter-channel level difference, and the inter-channel coherence, can be estimated from the covariance matrix of the analysis beamformers while the robust beamformers provide a high-quality signal for rendering. The synthesis beamformers are processed in the time-frequency domain with least-squares mixing and decorrelation to obtain the estimated target parametric properties. In other words, the technique proposes a signal-dependent parametric approach to spatial sound reproduction in order to combine the high spatial selectivity of the analysis patterns and the high signal quality of the broader patterns. The parametrisation accounts for any number of incoherent or mutually dependent sound sources, providing fundamental robustness for varying types of sound fields. Listening tests were organised to evaluate the performance of the system in the five-channel surround system. The proposed algorithm was shown to improve the perceptual aspects of loudspeaker rendering when compared to conventional signal-independent and signal-dependent systems under certain sound field conditions with respect to a reference signal.

6.4 Parametric binaural reproduction for compact microphone arrays

In Publication IV, the concept of enhancing a set of noise-robust beamformers that lack directional selectivity in a certain frequency range with optimal mixing parametric techniques is applied to headphone reproduction. In this case, the analysis beamforming weights are estimated with least-squares minimisation techniques and approximate the directional characteristics of a set of HRTFs with high accuracy. The weights for the synthesis beamformers are estimated with highly regularised least-squares minimisation to ensure a set of high-quality signals. These signals are then perceptually enhanced in the time-frequency domain with

a set of parameters estimated from the covariance matrix of the analysis beamformers. A simulation is provided to show the potential improvement in binaural reproduction for compact microphone arrays when compared to conventional beamforming techniques.

6.5 Direction-of-arrival estimation with histogram analysis of steered-response beamformers

Publications V considers the problem of DOA estimation, which is a required parameter for algorithms such as the ones proposed in Publications I and II. Steered-response power beamformers can provide highly accurate DOA estimation. However, their performance is degraded in reverberant environments. In this publication, these DOA estimators are enhanced by using 2-D histogram analysis. Four types of beamformers, all of them axis symmetric and signal-independent, have been studied in the spherical harmonic domain. The histogram analysis of the DOA estimates from the steered beamformers was shown to provide smoother power maps and to be able to reveal the DOAs of multiple sound sources. An improvement is shown in terms of DOA accuracy for a different number of sources with different angular separation and under different signal-to-noise conditions.

6.6 Direction-of-arrival estimation with histogram analysis of spatially constrained active intensity vectors

Publication VI studies DOA estimation using histogram analysis of higher-order active intensity vector DOA estimates. In contrast to other DOA estimators, such as steered-response power, active intensity is especially attractive due to its very low computational cost. In publication VI, the use of a spatially-constrained active intensity is utilised to minimise the effect of interfering sources that reside outside the analysis region. The instantaneous DOA from the active intensity are post processed with 1-D histogram analysis. DOAs of multiple sources can be then extracted from the histogram. A seven-channel mobile-like array prototype is being constructed to evaluate the performance of the algorithm in reverber-

ant conditions. The advantage of the spatially constrained active intensity vector is demonstrated in the sound field conditions of multiple non-coherent sources when a coherent source is present outside the analysis region.

6.7 Acoustic camera based on spatial parameters

In Publication VII, a real-time acoustic camera is implemented. The acoustic camera utilises in principle any microphone array input. As part of the software, the microphone input signals are initially transformed in the spherical harmonic domain and afterwards different types of power maps and DOA estimators are visualised. The power maps are based on signal-independent axis-symmetric beamformers as well as adaptive beamformers such as the MVDR. Subspace estimators such as MUSIC with the direct-path-dominance tests is also implemented. A novel beamforming-based approach is proposed that builds on the CroPaC spatial filter for capturing and analysing a sound field by estimating a probability-like parameter of sounds at specific locations. It is based on determining the correlation between two coincident beamformers. Conventional beamforming-based power maps perform poorly in reverberant and noisy conditions, due to the side-lobes of the beams. An algorithm to suppress side-lobes is shown, based on the product of multiple CroPaC beams is proposed. Illustrative power map examples are provided in highly reverberant spaces utilising a spherical microphone array.

7. Conclusions

Technologies in the field of parametric spatial audio processing are discussed in this thesis. The findings support the feasibility of compact microphone arrays in beamforming and spatial audio reproduction. The thesis provides contributions that are related to three aspects of microphone array processing techniques: parametric directional filtering, perceptually-motivated spatial sound reproduction, and DOA estimation.

One of the common issues in existing algorithms for spatial sound processing that are based on beamforming with compact microphone arrays is the actual design of the beamformers. Robust designs will result to beam patterns with frequency varying directivity and typically wider at low frequencies, while constant-directivity designs will result to noise amplification at low frequencies. In directional filtering applications this will cause a spectral imbalance for the direction of interest while in spatial sound reproduction high inter-channel coherence in the output signals. In this thesis we demonstrated that combining different types of beamformers can be beneficial in spatial sound processing applications.

In parameter-based directional filtering, it has been demonstrated that higher noise reduction can be achieved by estimating a probability-like parameter that indicates whether there is presence of a target signal at a specific direction. The calculation of this parameter is based on the cross-spectrum of two types beamformers and is then applied as a post filter at the output of another beamformer or single microphone to provide additional directional noise attenuation. In parametric spatial sound reproduction, high sound quality beamformers can be directionally enhanced using adaptive optimal mixing techniques with sound field parameters estimated from a set of beamformers with high directional selectivity. The

Conclusions

algorithms have been compared with the state-of-the-art and an improvement is shown using instrumental measured and listening tests.

References

- [1] Ville Pulkki and Matti Karjalainen. *Communication acoustics: an introduction to speech, audio and psychoacoustics*. John Wiley & Sons, 2015.
- [2] Ville Pulkki, Symeon Delikaris-Manias, and Archontis Politis. *Parametric time-frequency domain spatial audio*. Wiley, 2018.
- [3] Heinrich Kuttruff. *Room acoustics*. CRC Press, 2009.
- [4] Michael Barron. *Auditorium acoustics and architectural design*. Routledge, 2009.
- [5] *Acoustics — Measurement of the reverberation time of rooms with reference to other acoustic parameters*. ISO 3382:1997, 1997.
- [6] Frank Fahy. *Sound intensity*. CRC Press, 2002.
- [7] Finn Jacobsen. The diffuse sound field. Technical Report 27, The Acoustics Laboratory, Technical University of Denmark (DTU), Lyngby, Denmark, 1979.
- [8] John Eargle. *The Microphone Book: From mono to stereo to surround-a guide to microphone design and application*. CRC Press, 2012.
- [9] Hai Morgenstern, Boaz Rafaely, and Markus Noisternig. Design framework for spherical microphone and loudspeaker arrays in a multiple-input multiple-output system. *The Journal of the Acoustical Society of America*, 141(3):2024–2038, 2017.
- [10] Bradford N. Gover, James G. Ryan, and Michael R. Stinson. Microphone array measurement system for analysis of directional and spatial variations of sound fields. *The Journal of the Acoustical Society of America*, 112(5):1980–1991, 2002.
- [11] Philippe-Aubert Gauthier, Éric Chambatte, Cédric Camier, Yann Pasco, and Alain Berry. Beamforming regularization, scaling matrices and inverse problems for sound field extrapolation and characterization: Part i-theory. In *Audio Engineering Society Convention 131*. Audio Engineering Society, 2011.

- [12] Yoshio Yamasaki and Takeshi Itow. Measurement of spatial information in sound fields by closely located four point microphone method. *Journal of the Acoustical Society of Japan (E)*, 10(2):101–110, 1989.
- [13] Joseph Hector Dibiase, Harvey F. Silverman, and Michael S. Brandstein. Robust Localization in Reverberant Rooms. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 8, pages 157–180. Springer-Verlag, 2001.
- [14] Dima Khaykin and Boaz Rafaely. Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pages 221–224. IEEE, 2009.
- [15] Despoina Pavlidi, Anthony Griffin, Matthieu Puigt, and Athanasios Mouchtaris. Real-time multiple sound source localization and counting using a circular microphone array. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2193–2206, 2013.
- [16] Joonas Nikunen and Tuomas Virtanen. Direction of arrival based spatial covariance model for blind sound source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(3):727–739, 2014.
- [17] Barry D Van Veen and Kevin M Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP magazine*, 5(2):4–24, 1988.
- [18] Jacob Benesty, Jingdong Chen, and Yiteng Huang. *Microphone array signal processing*. Springer Topics in Signal Processing. Springer, 2008.
- [19] Peter Fellgett. Ambisonics. Part one: General system description. *Studio Sound*, 17(8):20–22, 1975.
- [20] Michael A Gerzon. Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, 21(1):2–10, 1973.
- [21] Jérôme Daniel. *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. PhD thesis, University of Paris VI, Paris, France, 2000.
- [22] Sébastien Moreau, Jérôme Daniel, and Stéphanie Bertet. 3d sound field recording with higher order ambisonics—objective measurements and validation of a 4th order spherical microphone. In *Audio Engineering Society Convention 120*, pages 20–23, 2006.
- [23] Ville Pulkki. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516, 2007.
- [24] Jacob Benesty and Chen Jingdong. *Study and design of differential microphone arrays*, volume 6. Springer Science & Business Media, 2012.
- [25] Jerome Daniel and Nicolas Epain. Improving spherical microphone arrays. In *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

- [26] Heinz Teutsch. *Modal array signal processing: principles and applications of acoustic wavefield decomposition*, volume 348. Springer, 2007.
- [27] Boaz Rafaely. *Fundamentals of Spherical Array Processing*, volume 8. Springer, 2015.
- [28] Angelo Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio Engineering Society Convention 108*. Audio Engineering Society, 2000.
- [29] Brian CJ Moore. *An introduction to the psychology of hearing*, volume 4. Academic press San Diego, 2003.
- [30] Brian R Glasberg and Brian CJ Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1):103–138, 1990.
- [31] Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997.
- [32] William M Hartmann. Localization of sound in rooms. *The Journal of the Acoustical Society of America*, 74(5):1380–1391, 1983.
- [33] Juha Vilkkamo. *Perceptually motivated time-frequency processing of spatial audio*. Doctoral dissertation, School of Electrical Engineering, Department of Signal Processing and Acoustics, 2014.
- [34] Thibaud Necciari, Peter Balazs, Nicki Holighaus, and Peter L Søndergaard. The erblet transform: An auditory-based time-frequency representation with perfect reconstruction. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 498–502. IEEE, 2013.
- [35] Jürgen Herre, Kristofer Kjörning, Jeroen Breebaart, Christof Faller, Sascha Disch, Heiko Purnhagen, Jeroen Koppens, Johannes Hilpert, Jonas Rödén, Werner Oomen, et al. Mpeg surround-the iso/mpeg standard for efficient and compatible multichannel audio coding. *Journal of the Audio Engineering Society*, 56(11):932–955, 2008.
- [36] Juha Vilkkamo and Symeon Delikaris-Manias. Perceptual reproduction of spatial sound using loudspeaker-signal-domain parametrization. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 23(10):1660–1669, Oct 2015.
- [37] Constantin Paleologu, Jacob Benesty, and Silviu Ciochină. Variable step-size adaptive filters for echo cancellation. In *Speech Processing in Modern Communication*, pages 89–125. Springer, 2010.
- [38] Michael S Brandstein and Harvey F Silverman. A practical methodology for speech source localization with microphone arrays. *Computer Speech Language*, 11(2):91–126, 1997.
- [39] Juha Merimaa. Applications of a 3-d microphone array. In *Audio Engineering Society Convention 112*. Audio Engineering Society, 2002.

- [40] Jukka Ahonen and Ville Pulkki. Speech intelligibility in teleconference application of directional audio coding. In *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*. Audio Engineering Society, 2010.
- [41] Jürgen Herre, Cornelia Falch, Dirk Mahane, Giovanni Del Galdo, Markus Kallinger, and Oliver Thiergart. Interactive teleconferencing combining spatial audio object coding and dirac technology. *Journal of the Audio Engineering Society*, 59(12):924–935, 2012.
- [42] Israel Cohen, Jacob Benesty, and Sharon Gannot. *Speech processing in modern communication: challenges and perspectives*, volume 3. Springer Science & Business Media, 2009.
- [43] Jonas Braasch. A loudspeaker-based 3d sound projection using virtual microphone control (vimic). In *Audio Engineering Society Convention 118*. Audio Engineering Society, 2005.
- [44] Joerg Bitzer and K Uwe Simmer. Superdirective microphone arrays. In *Microphone arrays*, pages 19–38. Springer, 2001.
- [45] Angelo Farina, Andrea Capra, Lorenzo Chiesi, and Leonardo Scopece. A spherical microphone array for synthesizing virtual directive microphones in live broadcasting and in post production. In *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*. Audio Engineering Society, 2010.
- [46] Lucas C Parra. Least squares frequency-invariant beamforming. In *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pages 102–105. IEEE, 2005.
- [47] Zhan Shi and Zhenghe Feng. A new array pattern synthesis algorithm using the two-step least-squares method. *IEEE signal processing letters*, 12(3):250–253, 2005.
- [48] Harry L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I*. John Wiley & Sons, Inc., New York, USA, September 2001.
- [49] Eugen Rasumow, Martin Hansen, Steven van de Par, Dirk Püschel, Volker Mellert, Simon Doclo, and Matthias Blau. Regularization approaches for synthesizing hrtf directivity patterns. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(2):215–225, 2016.
- [50] Joshua Atkins. Robust beamforming and steering of arbitrary beam patterns using spherical arrays. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pages 237–240. IEEE, 2011.
- [51] Maximo Cobos, Sascha Spors, Jens Ahrens, and Jose J Lopez. On the use of small microphone arrays for wave field synthesis auralization. In *Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*. Audio Engineering Society, 2012.

- [52] Franz Zotter and Matthias Frank. All-round ambisonic panning and decoding. *Journal of the audio engineering society*, 60(10):807–820, 2012.
- [53] Oliver Thiergart, Maja Taseska, and Emanuël AP Habets. An informed mmse filter based on multiple instantaneous direction-of-arrival estimates. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, pages 1–5. IEEE, 2013.
- [54] Oliver Thiergart and Emanuël AP Habets. An informed lcmv filter based on multiple instantaneous direction-of-arrival estimates. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 659–663. IEEE, 2013.
- [55] Oliver Thiergart and Emanuel AP Habets. Extracting reverberant sound using a linearly constrained minimum variance spatial filter. *IEEE Signal Processing Letters*, 21(5):630–634, 2014.
- [56] Oliver Thiergart, Maja Taseska, and Emanuël AP Habets. An informed parametric spatial filter based on instantaneous direction-of-arrival estimates. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(12):2182–2196, 2014.
- [57] Nobutaka Ito, Nobutaka Ono, Emmanuel Vincent, and Shigeki Sagayama. Designing the wiener post-filter for diffuse noise suppression using imaginary parts of inter-channel cross-spectra. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 2818–2821. IEEE, 2010.
- [58] K. Uwe Simmer, Joerg Bitzer, and Claude Marro. *Microphone arrays: signal processing techniques and applications*, chapter Post-Filtering Techniques, pages 39–60. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.
- [59] Pasi Pertilä and Joonas Nikunen. Microphone array post-filtering using supervised machine learning for speech enhancement. In *INTER-SPEECH*, pages 2675–2679, 2014.
- [60] Pasi Pertilä and Joonas Nikunen. Distant speech separation using predicted time–frequency masks from spatial features. *Speech Communication*, 68:97–106, 2015.
- [61] Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, and Kazunobu Kondo. Musical-Noise Analysis in Methods of Integrating Microphone Array and Spectral Subtraction Based on Higher-Order Statistics. *EURASIP Journal on Advances in Signal Processing*, 2010(1):431347, 2010.
- [62] JB Allen, DA Berkley, and J Blauert. Multimicrophone signal-processing technique to remove room reverberation from speech signals. *The Journal of the Acoustical Society of America*, 62(4):912–915, 1977.
- [63] Iain A McCowan and Hervé Bouchard. Microphone array post-filter based on noise field coherence. *Speech and Audio Processing, IEEE Transactions on*, 11(6):709–716, 2003.

- [64] Rainer Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 2578–2581. IEEE, 1988.
- [65] Simon Doclo, Walter Kellermann, Shoji Makino, and Sven Erik Nordholm. Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones. *IEEE Signal Processing Magazine*, 32(2):18–30, 2015.
- [66] Markus Kallinger, Henning Ochsenfeld, Giovanni Del Galdo, Fabian Kuech, Dirk Mahne, Richard Schultz-Amling, and Oliver Thiergart. A spatial filtering approach for directional audio coding. In *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.
- [67] Christof Faller. Modifying the directional responses of a coincident pair of microphones by postprocessing. *Journal of the Audio Engineering Society*, 56(10):810–822, 2008.
- [68] Christof Faller. A highly directive 2-capsule based microphone. In *Audio Engineering Society Convention 123*. Audio Engineering Society, 2007.
- [69] Boaz Rafaely. Analysis and design of spherical microphone arrays. *IEEE Transactions on Speech and Audio Processing*, 13(1):135–143, 2005.
- [70] Symeon Delikaris-Manias, Juha Vilkamo, and Ville Pulkki. Parametric binaural rendering utilizing compact microphone arrays. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 629–633, April 2015.
- [71] Earl G Williams. *Fourier acoustics: sound radiation and nearfield acoustical holography*. Academic press, 1999.
- [72] Thushara D Abhayapala. Generalized framework for spherical microphone arrays: Spatial and frequency decomposition. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 5268–5271. IEEE, 2008.
- [73] David L Alon, Jonathan Sheaffer, and Boaz Rafaely. Robust plane-wave decomposition of spherical microphone array recordings for binaural sound reproduction. *The Journal of the Acoustical Society of America*, 138(3):1925–1926, 2015.
- [74] Franz Zotter, Hannes Pomberger, and Markus Noisternig. Energy-preserving ambisonic decoding. *Acta Acustica united with Acustica*, 98(1):37–47, 2012.
- [75] Bo Wang, Yanping Zhao, and Juanjuan Liu. Mixed-order MUSIC algorithm for localization of far-field and near-field sources. *IEEE Signal Processing Letters*, 20(4):311–314, April 2013.

- [76] Or Nadiri and Boaz Rafaely. Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(10):1494–1505, 2014.
- [77] Huseyin Hacihabiboglu. Theoretical analysis of open spherical microphone arrays for acoustic intensity measurements. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(2):465–476, 2014.
- [78] Banu Günel and H Hacihabiboglu. Sound source localization: Conventional methods and intensity vector direction exploitation. In Wenwu Wang, editor, *Machine Audition: Principles, Algorithms and Systems*, chapter 6, pages 126–161. IGI Global, 2011.
- [79] Daniel P Jarrett, Emanuël AP Habets, and Patrick A Naylor. 3D source localization in the spherical harmonic domain using a pseudointensity vector. In *18th European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark, 2010.
- [80] Despoina Pavlidi, Symeon Delikaris-Manias, Ville Pulkki, and Athanasios Mouchtaris. 3d doa estimation of multiple sound sources based on spatially constrained beamforming driven by intensity vectors. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 96–100. IEEE, 2016.
- [81] Christine Evers, Alastair H Moore, Patrick Naylor, et al. Multiple source localisation in the spherical harmonic domain. In *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on*, pages 258–262. IEEE, 2014.
- [82] Despoina Pavlidi, Symeon Delikaris-Manias, Ville Pulkki, and Athanasios Mouchtaris. 3d localization of multiple sound sources with intensity vector estimates in single source zones. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 1556–1560. IEEE, 2015.
- [83] Alastair H Moore, Christine Evers, Patrick A Naylor, David L Alon, and Boaz Rafaely. Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test. In *23rd European Signal Processing Conf. (EUSIPCO)*, Nice, Italy, 2015.
- [84] Sakari Tervo. Direction estimation based on sound intensity vectors. In *17th European Signal Processing Conference (EUSIPCO)*, pages 700–704, Glasgow, UK, 2009.
- [85] Dovid Levin, Emanuël a P Habets, and Sharon Gannot. On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields. *The Journal of the Acoustical Society of America*, 128(4):1800–11, October 2010.
- [86] Haohai Sun, Heinz Teutsch, Edwin Mabande, and Walter Kellermann. Robust localization of multiple sources in reverberant environments using

- eb-esprit with spherical microphone arrays. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 117–120. IEEE, 2011.
- [87] Symeon Delikaris-Manias, Despoina Pavlidi, Ville Pulkki, and Athanasios Mouchtaris. 3d localization of multiple audio sources utilizing 2d doa histograms. In *Signal Processing Conference (EUSIPCO), 2016 24th European*, pages 1473–1477. IEEE, 2016.
- [88] Xuan Li, Shefeng Yan, Xiaochuan Ma, and Chaohuan Hou. Spherical harmonics music versus conventional music. *Applied Acoustics*, 72(9):646–652, 2011.
- [89] Juha Merimaa and Ville Pulkki. Spatial impulse response rendering i: Analysis and synthesis. *Journal of the Audio Engineering Society*, 53(12):1115–1127, 2005.
- [90] Juha Merimaa. *Analysis, synthesis, and perception of spatial sound: binaural localization modeling and multichannel loudspeaker reproduction*. PhD thesis, Laboratory of Acoustics and Audio Signal Processing, Espoo, Finland, 2006.
- [91] Ville Pulkki, Archontis Politis, Giovanni Del Galdo, and Achim Kuntz. Parametric spatial audio reproduction with higher-order B-Format microphone input. In *Audio Engineering Society Convention 134*, May 2013.
- [92] Audun Solvang. Spectral impairment of two-dimensional higher order Ambisonics. *Journal of the Audio Engineering Society*, 56(4):267–279, 2008.
- [93] Peter Stitt, Stéphanie Bertet, and Maarten van Walstijn. Off-centre localisation performance of ambisonics and HOA for large and small loudspeaker array radii. *Acta Acustica united with Acustica*, 100(5):937–944, 2014.
- [94] Liu Yang and Xie Bosun. Subjective evaluation on the timbre of horizontal ambisonics reproduction. In *Audio, Language and Image Processing (ICALIP), 2014 International Conference on*, pages 11–15. IEEE, 2014.
- [95] Günther Theile. Natural 5.1 music recording based on psychoacoustic principals. In *19th Int. Conf. of the AES: Surround Sound-Techniques, Technology, and Perception*, Bavaria, Germany, 2001.
- [96] Michael Williams. Multichannel sound recording using 3, 4 and 5 channels arrays for front sound stage coverage. In *Audio Engineering Society Convention 117*. Audio Engineering Society, 2004.
- [97] Francis Rumsey. *Spatial audio*. CRC Press, 2012.
- [98] Ville Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466, 1997.
- [99] Mark A Poletti. Three-dimensional surround sound systems based on spherical harmonics. *Journal of the Audio Engineering Society*, 53(11):1004–1025, 2005.

- [100] Yoomi Hur, Jonathan S Abel, Young-Cheol Park, and Dae Hee Youn. Techniques for synthetic reconfiguration of microphone arrays. *Journal of the Audio Engineering Society*, 59(6):404–418, 2011.
- [101] Angelo Farina, Alberto Amendola, Lorenzo Chiesi, Andrea Capra, and Simone Campanini. Spatial pcm sampling: A new method for sound recording and playback. In *Audio Engineering Society Conference: 52nd International Conference: Sound Field Control-Engineering and Perception*. Audio Engineering Society, 2013.
- [102] Jiashu Chen, Barry D Van Veen, and Kurt E Hecox. External ear transfer function modeling: A beamforming approach. *The Journal of the Acoustical Society of America*, 92(4):1933–1944, 1992.
- [103] Shuichi Sakamoto, Jun’ichi Kodama, Satoshi Hongo, Takuma Okamoto, Yukio Iwaya, and Yôiti Suzuki. A 3D sound-space recording system using spherical microphone array with 252ch microphones. In *20th International Congress on Acoustics (ICA)*, Sydney, Australia, 2010.
- [104] Jukka Ahonen. *Microphone front-ends for spatial sound analysis and synthesis using directional audio coding*. Doctoral dissertation, School of Electrical Engineering, Department of Signal Processing and Acoustics, Espoo, Finland, 2013.
- [105] Mikko-Ville Laitinen. *Techniques for versatile spatial-audio reproduction in time-frequency domain*. Doctoral dissertation, School of Electrical Engineering, Department of Signal Processing and Acoustics, Espoo, Finland, 2014.
- [106] Marko Takanen, Olli Santala, and Ville Pulkki. Binaural assessment of parametrically coded spatial audio signals. In *The technology of binaural listening*, pages 333–358. Springer, 2013.
- [107] Karlheinz Brandenburg. Mp3 and aac explained. In *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*. Audio Engineering Society, 1999.
- [108] Huseyin Hacihabiboglu, Enzo De Sena, Zoran Cvetkovic, James. D Johnston, and Julius O. Smith III. Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics. *IEEE Signal Processing Magazine*, 34(3):36–54, May 2017.
- [109] Christof Faller. Parametric multichannel audio coding: synthesis of coherence cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):299–310, 2006.
- [110] Frank Baumgarte and Christof Faller. Binaural cue coding-part I: Psychoacoustic fundamentals and design principles. *IEEE Transactions on Speech and Audio Processing*, 11(6):509–519, 2003.
- [111] Christof Faller and Frank Baumgarte. Binaural cue coding-part II: Schemes and applications. *IEEE Transactions on Speech and Audio Processing*, 11(6):520–531, 2003.

- [112] Jeroen Breebaart, Steven van de Par, Armin Kohlrausch, and Erik Schuijers. Parametric coding of stereo audio. *EURASIP Journal on Applied Signal Processing*, 2005:1305–1322, 2005.
- [113] Maximo Cobos, Jose J. Lopez, and Sascha Spors. A sparsity-based Approach to 3D binaural sound synthesis using time-frequency array processing. *EURASIP Journal on Advances in Signal Processing*, 2010:1–13, 2010.
- [114] Juan José Burred and Thomas Sikora. On the use of auditory representations for sparsity-based sound source separation. In *Information, Communications and Signal Processing, 2005 Fifth International Conference on*, pages 1466–1470. IEEE, 2005.
- [115] Natasha Barrett and Svein Berge. A new method for b-format to binaural transcoding. In *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*. Audio Engineering Society, 2010.
- [116] Oliver Thiergart and Emanuël AP Habets. Robust direction-of-arrival estimation of two simultaneous plane waves from a b-format signal. In *Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of*, pages 1–5. IEEE, 2012.
- [117] Anastasios Alexandridis, Anthony Griffin, and Athanasios Mouchtaris. Capturing and reproducing spatial audio based on a circular microphone array. *Journal of Electrical and Computer Engineering*, 2013:1–16, 2013.
- [118] Jon Ander Beracoechea, Javier Casajus, Lino García, Luis Ortiz, and Soledad Torres-Guijarro. Implementation of immersive audio applications using robust adaptive beamforming and wave field synthesis. In *Audio Engineering Society Convention 120*. Audio Engineering Society, 2006.
- [119] Ville Pulkki and Juha Merimaa. Spatial impulse response rendering ii: Reproduction of diffuse sound and listening tests. *Journal of the Audio Engineering Society*, 54(1/2):3–20, 2006.
- [120] Jukka Ahonen, Giovanni Del Galdo, Fabian Kuech, and Ville Pulkki. Directional analysis with microphone array mounted on rigid cylinder for directional audio coding. *Journal of the Audio Engineering Society*, 60(5):311–324, 2012.
- [121] Jukka Ahonen, Giovanni Del Galdo, Markus Kallinger, Fabian Küch, Ville Pulkki, and Richard Schultz-Amling. Analysis and adjustment of planar microphone arrays for application in directional audio coding. In *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.
- [122] Mikko-Ville Laitinen, Fabian Küch, and Ville Pulkki. Using spaced microphones with directional audio coding. In *Audio Engineering Society Convention 130*, May 2011.
- [123] Oliver Thiergart, Michael Kratschmer, Markus Kallinger, and Giovanni Del Galdo. Parameter Estimation in Directional Audio Coding Using Linear Microphone Arrays. *New York*, 2011.

- [124] Jukka Ahonen and Ville Pulkki. Diffuseness estimation using temporal variation of intensity vectors. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pages 285–288. IEEE, 2009.
- [125] Giovanni Del Galdo, Maja Taseska, Oliver Thiergart, Jukka Ahonen, and Ville Pulkki. The diffuse sound field in energetic analysis. *The Journal of the Acoustical Society of America*, 131(3):2141–2151, 2012.
- [126] Jukka Ahonen and Ville Pulkki. Broadband direction estimation method utilizing combined pressure and energy gradients from optimized microphone array. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 97–100. IEEE, 2011.
- [127] O. Thiergart, G. D. Galdo, and E. A. P. Habets. Diffuseness estimation with high temporal resolution via spatial coherence between virtual first-order microphones. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 217–220, Oct 2011.
- [128] Juha Vilkkamo, Tapio Lokki, and Ville Pulkki. Directional audio coding: Virtual microphone-based synthesis and subjective evaluation. *Journal of the Audio Engineering Society*, 57(9):709–724, 2009.
- [129] Juha Vilkkamo and Ville Pulkki. Minimization of decorrelator artifacts in directional audio coding by covariance domain rendering. *Journal of the Audio Engineering Society*, 61(9):637–646, 2013.
- [130] Archontis Politis. *Microphone array processing for parametric spatial audio techniques*. Doctoral dissertation, School of Electrical Engineering, Department of Signal Processing and Acoustics, Espoo, Finland, 2016.
- [131] Jeroen Breebaart, Jonas Engdegård, Cornelia Falch, Oliver Hellmuth, Johannes Hilpert, Andreas Hoelzer, Jeroen Koppens, Werner Oomen, Barbara Resch, Erik Schuijers, et al. Spatial audio object coding (saoc)-the upcoming mpeg standard on parametric object based audio coding. In *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.
- [132] Leonid Terentiev, Cornelia Falch, Oliver Hellmuth, Johannes Hilpert, Werner Oomen, Jonas Engdegård, and Mundt Harald. Saoc for gaming—the upcoming mpeg standard on parametric object based audio coding. In *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Audio Engineering Society, 2009.
- [133] Jean-Marc Jot. Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces. *Multimedia systems*, 7(1):55–69, 1999.
- [134] Manuel Briand, Nadine Martin, and David Virette. Parametric representation of multichannel audio based on principal component analysis. In *Audio Engineering Society Convention 120*. Audio Engineering Society, 2006.

References

- [135] Joonas Nikunen, Tuomas Virtanen, and Miikka Vilermo. Multichannel audio upmixing based on non-negative tensor factorization representation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2011.

Spatial hearing provides us with information about our surroundings: the location of multiple concurrent sound sources along with their corresponding distance, information about the space they reside in and their content. Parametric spatial sound technologies are multichannel signal processing algorithms that aim to capture, analyse and render a complex sound scene for a human listener in order that he or she obtains an understanding of the environment or has the auditory experience of being there.

This dissertation focuses on the development of parametric spatial audio technologies using prototype and commercial compact microphone arrays. Compact arrays are of special interest since they can be adapted to fit in portable devices, opening the possibility of exploiting the potential of immersive spatial audio algorithm in our daily lives. The findings of this research are in the following three areas of spatial audio processing: directional filtering, spatial audio reproduction, and direction of arrival estimation.



ISBN 978-952-60-7661-4 (printed)

ISBN 978-952-60-7660-7 (pdf)

ISSN-L 1799-4934

ISSN 1799-4934 (printed)

ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Department of Signal Processing and Acoustics
www.aalto.fi

**BUSINESS +
ECONOMY**

**ART +
DESIGN +
ARCHITECTURE**

**SCIENCE +
TECHNOLOGY**

CROSSOVER

**DOCTORAL
DISSERTATIONS**