Department of Signal Processing and Acoustics

Parametric reproduction of microphone array recordings

Leo McCormack

Aalto University

DOCTORAL THESES

Aalto University publication series
DOCTORAL THESES 60/2023

Parametric reproduction of microphone array recordings

Leo McCormack

A doctoral thesis 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 lecture hall U142 U4 Otakaari 1 of the school on 11th of May 2023 at 12:15.

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

Supervising professor
Prof. Ville Pulkki, Aalto University, Finland

Thesis advisors
Prof. Archontis Politis, Tampere University, Finland
Prof. Ville Pulkki, Aalto University, Finland

Preliminary examiners
Dr. Nicolas Epain, b<>com, France
Dr. Oliver Thiergart, Fraunhofer IIS, Germany

Opponent
Prof. Franz Zotter, Universität für Musik und darstellende Kunst Graz, Austria

Aalto University publication series
DOCTORAL THESES 60/2023

ISBN 978-952-64-1244-3 (printed)
ISBN 978-952-64-1245-0 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-1245-0

Unigrafia Oy
Helsinki 2023

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

Author
Leo McCormack

Name of the doctoral thesis
Parametric reproduction of microphone array recordings

Publisher School of Electrical Engineering

Unit Department of Signal Processing and Acoustics

Series Aalto University publication series DOCTORAL THESES 60/2023

Field of research Acoustics and Audio Signal Processing

Manuscript submitted 17 April 2023 **Date of the defence** 11 May 2023

Permission for public defence granted (date) 6 April 2023 **Language** English

☐ **Monograph** ☒ **Article thesis** ☐ **Essay thesis**

Abstract

This thesis encloses five publications which describe technologies for recording, analysing, manipulating, and reproducing spatial sound scenes, which confront many of the challenges associated with the development of systems capable of delivering high quality audio within virtual reality and augmented hearing contexts. The technologies detailed herein operate based upon microphone array signals, which have been transformed into the time-frequency domain. Through the adoption of an assumed sound-field model, an input sound scene may be parameterised and decomposed, which permits the optional manipulation and subsequent reproduction of the sound scene over an arbitrary playback setup. This type of processing often leads to a high degree of playback flexibility and perceived spatial accuracy, which would otherwise be unattainable when using signal-independent and non-parametric alternatives. The first contribution of this thesis concerns the parameterisation and rendering of microphone array room impulse responses, such that the spatial characteristics of a measured space may be imparted onto a monophonic input signal and reproduced over a target loudspeaker setup. The second contribution explores a parametric method for converting microphone array signals into the popular Ambisonics format, while placing specific emphasis on the use of microphone arrays that are mounted onto irregular/non-spherical geometries; such as head-worn devices, which may find application within future augmented reality contexts. The third contribution also concerns a head-worn microphone array, but instead utilised microphones that are sensitive to ultrasonic frequencies. The intention is for ultrasonic sound sources to be captured by the array and then down pitch-shifted to the audible range, while being spatialised in the same direction that the sound arrived from. A number of spatial audio effects and sound-field modification tools were then explored in the fourth contribution, which operate based upon Ambisonic signals as input and involve the use of a parametric rendering framework. The final contribution concerns the use of a distributed arrangement of multiple Ambisonic receivers, which may be used to capture the sound scene from multiple perspectives. Subsequent analysis and decomposition of the sound scene, into its individual components, enables reproduction at different positions; thus, allowing a listener to navigate through the recorded sound scene.

Keywords spatial audio, array signal processing

ISBN (printed) 978-952-64-1244-3 **ISBN (pdf)** 978-952-64-1245-0

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2023

Pages 120 **urn** http://urn.fi/URN:ISBN:978-952-64-1245-0

Preface

My story begins as an infant in Flemish Brabant, Belgium, in a village aptly named, Tremolo. My rocking chair is sat atop of my father's old upright piano, which is situated in the living room of my early childhood home. Gentle bespoke lullabies are being sung to me, in an effort to soothe and return me to a resting state. Fortunately, some combination of the piano vibrations and/or gentle melodies is oftentimes successful in this endeavour, and thus this ritual continued throughout much of early childhood - until I started joining in rather discordantly...

Upon my return to the UK, now aged almost seven, I spent many of my primary and secondary school years being infatuated with musical instruments; such as my father's upright piano and electronic keyboard, which had several other sounds and voices to play around with. I also became involved with a couple of high school rock/metal bands, which meant adding a synthesiser, a "couple" of guitars, amplifiers and pedals to my musical collection.

In the final years of secondary school, I began to gravitate towards recording and mixing my formative (and oftentimes questionable) musical creations. I was therefore overjoyed when my music teacher, Mr. Ralphson, informed me that a music technology A-level course had been introduced at the high school; since this allowed me to obtain a formal qualification for my experimental musical nonsense (genius).

Upon completion of my secondary school education, I wished to continue down this recording and music production path. Therefore, I chose to study for a bachelor's degree in Music Technology and Audio Systems at the University of Huddersfield, UK. At some point after the first year of university studies, and much to my own surprise, I also began to pick up an interest in courses related to audio signal processing, programming, and psychoacoustics. My year long internship at Fraunhofer IIS, Erlangen, Germany, and the encouragement of my mentor at the time, Prof. Hyunkook Lee, solidified this shift in study direction. While I was still happy to create and mix my own music, I realised I had become more interested in learning about the inner-workings of the tools I was using.

It was also around this time that I was introduced to the world of spatial audio. My auditory system appeared to be no longer completely satisfied with mono and stereo mixes, and after returning to Huddersfield, I elected to pursue a final year project related to the (*vector-base-based*) vector-base amplitude panning method. This project didn't change the world, but it did serve as confirmation to myself that perhaps my spatial audio technology journey was only just beginning.

Soon after graduating from the University of Huddersfield, I was beginning a Master's in Acoustics and Audio Technology at Aalto University, Finland. Coming from a more musical background, while being surrounded by fellow students who were (mostly) electrical engineers, I perhaps had a bit of a tougher time working through some of the more maths-heavy aspects of the courses in the beginning. However, I believe that I developed a pretty good understanding of how these audio technologies could perhaps be integrated into tools and used in practice. Therefore, early on in the master's program, I approached Prof. Ville Pulkki and asked him whether I would be able to work on some spatial audio related projects in his research group. Having demonstrated that I could potentially be up to such a task (by solving a riddle, opening a puzzle box and commenting positively on certain dance moves), he accepted this request and introduced me to two of his PhD students, (now Dr.) Symeon Delikaris-Manias and (now Prof.) Archontis Politis. Over the course of many months, interleaved with other curriculum work, I proceeded to develop real-time audio plug-ins to demonstrate the practicality of their recent research endeavours. Admittedly, this side-hobby perhaps got a little bit out of hand in the years that followed... However, I learnt a lot during this time, and I am eternally grateful to these two individuals in particular, as they were very generous with lending me their time.

After graduating from the master's course, I somehow seamlessly segued into the PhD program, which has ultimately led to the writing of this heartfelt thesis preface, outlining my backstory.

This work was conducted at the Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland.

I would like to extend my eternal gratitude to supervisor Prof. Ville Pulkki and thesis advisor Prof. Archontis Politis, for their limitless patience and invaluable guidance over these past few years.

I would also like to thank my thesis pre-examiners, Dr. Oliver Thiergart and Dr. Nicolas Epain, for their prompt reviews and constructive comments.

I also wish to make it known that I had the great pleasure to work alongside many wonderful people and diligent researchers during my time at the Aalto Acoustics Lab; namely: Dr. Symeon Delikaris-Manias, Dr. Henri Pöntynen, Juhani Paasonen, Dr. Alessandro Altoè, Dr. Javier Gómez Bolaños, Ilkka Huhtakallio, Dr. Catarina Hiipakka, Christoph Hold, Pedro Lladó, Georg Götz, Stephan Wirler, Aleksi Öyry, Vasileios Bountourakis, Ricardo Falcon-Perez, Taeho Kim, Dr. Petteri Hyvärinen, Mădălina Anamaria Năstasă, Dr. Thomas McKenzie, Michael McCrea, Lauros Pajunen, Rapolas Daugintis, Abraham Ornelas, Dr. Nelli Salminen, Kabir Carter, Viktoria Korshunova, Kuura Parkkola, Bryn Louise, Alice Bourgain-Wilbal, Vilppu Pekkanen, Juuso Tolonen, Tommi Niemelä, Otso Rautama, Ossi Miikkulainen, Prof. Sebastian Schlecht, Prof. Tapio Lokki, Dr. Henna Tahvanainen, Dr. Julie Meyer, Dr. Raimundo Gonzalez, Dr. Antti Kuusinen, Dr. Jukka Pätynen, Nils Meyer-Kahlen, Jose Cucharero Moya, Laura McLeod, Dr. Otto Puomio, Riionheimo Janne, Arif Yürek, Meryem Jabrane, Frini Paschou, Prof. Vesa Välimäki, Dr. Fabián Esqueda, Leonardo Fierro, Dr. Karolina Prawda, Dr. Benoit Alary, Etienne Thuillier, Dr. Jussi Rämö, Eero-Pekka Damskägg, Craig Rollo, Eloi Moliner, Dimitrios Koutsaidis, Gloria Dal Santo, Jan Wilczek, Aaron Geldert, Otto Mikkonen, Dr. Juho Liski, Alec Wright, Jon Fagerström, Prof. Lauri Savioja, and Dr. Sebastian Prepeliţă.

I also greatly appreciated the opportunity to collaborate with members of other research groups; namely: Oliver Scheuregger and Dr. Marton Marschall from the Technical University of Denmark; Prof. Angelo Farina and Dr. Daniel Pinardi from the University of Parma, Italy; Dr. Despoina Pavlidi from the Foundation for Research and Technology-Hellas, Greece; and Prof. Simo Särkkä from Aalto University, Finland.

I would additionally like to thank my hosts Dr. Antti Eronen, Dr. David Lou Alon, and Dr. Zamir Ben-Hur and acknowledge all the other amazing researchers that I met while at Nokia and Meta Reality Labs Research.

Many thanks are also extended to my parents and brother for their support over the years, especially my father, for taking a keen interest in this academic journey of mine, and also for proofreading this thesis and associated publications. Lastly, I would like to thank my wonderful wife and colleague, Janani Fernandez, without whom this would not have been possible.

Helsinki, Finland, April 17, 2023,

Leo McCormack

Contents

Preface **1**

Contents **5**

List of Publications **7**

Author's Contribution **9**

1. Introduction **11**

2. Summary of Contributions **19**
2.1 Rendering higher-order Ambisonic room impulse responses 19
2.2 Ambisonic encoding of arbitrary microphone arrays 23
2.3 Superhuman spatial hearing for ultrasonic sound sources 25
2.4 Spatial audio effects and sound-field modifications 26
2.5 Six-degrees-of-freedom rendering of spatial sound scenes . 30

References **33**

Publications **39**

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 Leo McCormack, Ville Pulkki, Archontis Politis, Oliver Scheuregger and Marton Marschall. Higher-order spatial impulse response rendering: Investigating the perceived effects of spherical order, dedicated diffuse rendering, and frequency resolution. *Journal of the Audio Engineering Society (JAES)*, vol. 68, no. 5, pp. 338–354, May 2020.

II Leo McCormack, Archontis Politis, Raimundo Gonzalez, Tapio Lokki and Ville Pulkki. Parametric Ambisonic Encoding of Arbitrary Microphone Arrays. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 30, June 2022.

III Ville Pulkki, Leo McCormack and Raimundo Gonzalez. Superhuman spatial hearing technology for ultrasonic frequencies. *Scientific Reports*, 11, 11608, June 2021.

IV Leo McCormack, Archontis Politis and Ville Pulkki. Parametric Spatial Audio Effects Based on the Multi-Directional Decomposition of Ambisonic Sound Scenes. In *Proceedings of the 24th International Conference on Digital Audio Effects (DAFx20in21)*, September 2021.

V Leo McCormack, Archontis Politis, Thomas McKenzie, Christoph Hold and Ville Pulkki. Object-Based Six-Degrees-of-Freedom Rendering of Sound Scenes Captured with Multiple Ambisonic Receivers. *Journal of the Audio Engineering Society (JAES)*, vol. 70, no. 5, pp. 355-372, May 2022.

Author's Contribution

Publication I: "Higher-order spatial impulse response rendering: Investigating the perceived effects of spherical order, dedicated diffuse rendering, and frequency resolution"

The present, second and third authors developed the described algorithm. The formal perceptual listening test was conducted by the fourth and fifth authors. The manuscript was primarily written by the first and second authors, with contributions from all the other authors.

Publication II: "Parametric Ambisonic Encoding of Arbitrary Microphone Arrays"

The present author developed the described algorithms in collaboration with the second author, and conducted the objective and formal perceptual evaluations of proposed method in collaboration with the second and third authors. The majority of the text was written by the present author, with contributions from all of the other authors.

Publication III: "Superhuman spatial hearing technology for ultrasonic frequencies"

The present author developed the signal processing solution employed by the described ultrasonic recording device. The three authors contributed equally to writing the manuscript.

Publication IV: "Parametric Spatial Audio Effects Based on the Multi-Directional Decomposition of Ambisonic Sound Scenes"

The present and second author developed the proposed algorithms. The majority of the text was written by the present author, with contributions from the other two authors.

Publication V: "Object-Based Six-Degrees-of-Freedom Rendering of Sound Scenes Captured with Multiple Ambisonic Receivers"

The present author developed the proposed system in collaboration with the second and third authors, and evaluated the system in collaboration with the third and fourth authors. The sections describing the proposed analysis, synthesis, and the implementation of the method were written primarily by the first author. The background section was written primarily by the second author.

1. Introduction

Spatial audio technologies based upon microphone arrays form an integral part of the capture and reproduction of immersive audio experiences. In recent years, such technologies have found application in the dynamic rendering of recorded sound scenes both within virtual reality (VR) environments, and for the enhancement, modification, and rendering of spatial attributes of sound scenes associated with hearing aids and augmented reality (AR) contexts. Technologies are also available to impart the acoustical properties of measured spaces onto other musical signals, and also for the general capture, transmission, and playback of musical performances and other soundscapes.

In the absence of having detailed prior knowledge of the surrounding room geometry and the locations of sound sources, a single microphone is unable to capture and represent all of the spatial properties of a sound scene. Therefore, the simplest spatial audio pipelines involve the use of two microphones, with their respective signals delivered directly over headphones or two loudspeakers (typically) located in front of the listener. The relative positions and the directivity characteristics of these two microphones are then carefully chosen, such that the relationships between the recorded signals will deliver the appropriate perceptual cues [8] to the listener when they are played over the chosen reproduction setup. For many of the aforementioned applications, these appropriate perceptual cues refer to those required for the listeners' experience of the reproduced sound scene to be identical, or at least substantially similar, to how it would have been, had the listener instead been located at the recording point. However, while two microphones may often be easily integrated within existing pipelines, the resulting playback flexibility and/or perceived spatial accuracy will always be negatively impacted to some degree. For example, if the two microphones exhibit directivity patterns that exactly match the listener's two ears, which may be realised by placing the microphones within the listener's ear canals during recording, then a complete and spatially accurate reproduction of the original scene can be achieved. However, such an approach is contingent on the signals being reproduced

over headphones which are worn by the same listener, and the orientation of the sound scene remains fixed precisely to how it was during the recording itself; i.e., cannot be altered to account for a different orientation of the listener's head. Furthermore, if the recording is experienced by a different listener, then the perceived spatial accuracy will also be reduced to some degree; since the shape of a persons pinna and surrounding head and shoulder geometry, which influence ones spatial perception of the scene, is highly individualistic. Therefore, the portability and flexibility of such an approach is limited in practice.

Regarding traditional audio recording and production, where loudspeakers are also widely employed as the playback setup, preserving the original spatial properties of the scene is still possible to some degree, but this may be considered to be a more relaxed requirement. Here, the selection of the microphone positions and their directivities is also influenced by the creative preferences of the recording engineer. The one-to-one mapping of two microphone signals to the respective loudspeakers, which is collectively referred to as stereophony [9], also inherently limits and constrains the rendering within a narrow angular range of operation; as dictated by the loudspeaker spacing. These one-to-one channel mapping principles may, however, also be extended to multi-microphone and multi-loudspeaker setups, which can permit reproduction of the spatial properties of the recorded sound scene over a wider area surrounding the listener. The application of the appropriate binaural filters to each loudspeaker signal also enables convenient reproduction of the same sound scene over headphones. However, while stereophony is a well-studied technique, scaling up these recording principles to larger loudspeaker arrays, while trying to preserve the spatial properties of the captured space, becomes increasingly difficult and convoluted as more channels are added; since such setups usually require custom rigs and non-obvious optimisations of the microphone array directivities and their relative time differences. Another drawback of such approaches, is that there may not always be a clear solution for reproducing a microphone array recording intended for one particular loudspeaker setup over a different loudspeaker system. Therefore, the flexibility afforded by these so-called *channel-based* recording and playback approaches can be limited.

Today, delivery of sound scene recordings within the consumer space is still dominated by channel-based approaches. However, alternative formats and technologies targeting more flexible reproduction of microphone array recordings are receiving increased attention. This is largely due to the emergence of AR and VR devices, and growing public interest in the consumption of 360 degree immersive audio-visual media. In the case of AR and VR, audio playback is typically achieved via a pair of headphones. Although, it should be acknowledged that some installations may instead use large custom loudspeaker arrays to deliver the intended immersive

audio experience to the listener(s). Therefore, in the interest of flexibility, it would be desirable for a compact recording device to be able to capture the whole sound scene, and for it to then be reproduced over a wide range of different playback setups. Such a pipeline should also be conducted in a manner that ensures the delivery of the most salient perceptual cues to a diverse set of listeners, who should all broadly have the same immersive audio experience. In AR/VR contexts, the audio perception must also align with the corresponding visuals and, thus, accounting for listener head-rotations becomes a key requirement. Note that systems accounting for head-rotations, such that sound sources may remain anchored within the rendered virtual or augmented world, are commonly described as offering *three-degrees-of-freedom*.

Moving away from approaches relying on one-to-one channel mappings, there are alternative pipelines that jointly take into consideration the specifications of the microphone array and the specifications of the playback setup, in order to derive linear and time-invariant multi-channel filters to conduct the mapping of channels. Such methods aim to improve the practical flexibility of spatial audio capture and reproduction, in cases where the playback system has information regarding the microphone array specifications. The popular Ambisonics spatial audio format [19] improves this flexibility further, by first converting (or *encoding*) microphone array signals into intermediate signals which exhibit the directivities of spherical harmonics [52]. This Ambisonics format represents an efficient means of storing a sound scene over the full 360 degrees perspective, with the frequency- and direction-dependent spatial resolution of this representation being determined by the number of microphones in the array, their relative placement, and the material properties and construction of the mounting hardware. Additionally, the format can also facilitate inherently efficient sound-field rotations to account for different listener head orientations. There are then a number of algorithms available for the task of mapping (or *decoding*) these intermediate spherical harmonic signals to arbitrary playback setups [72]. The Ambisonics format may therefore be described as a highly convenient recording- and playback-agnostic means of storing, transmitting, and modifying sound scenes.

It is also highlighted that, in addition to processing microphone array recordings, any established linear Ambisonics rendering pipeline may also be applied within the context of rendering microphone array room impulse responses. In this case the task is to render the input Ambisonic room impulse responses in an appropriate manner, such that they instead correspond to the target playback setup. By subsequently convolving a monophonic input signal with this rendered response, the signal may be reproduced over the playback setup, while retaining the spatial characteristics of the captured space which have been imparted onto it. Such processing may be useful for the perceptual evaluation and analysis of

rooms, since it allows, for example, for different concert halls to be compared side-by-side in a controlled setting.

On the surface, therefore, the Ambisonics format, accompanied by well-established linear rendering pipelines [72], could be viewed as the ideal solution for catering to all VR/AR and immersive audio applications. In practice, however, the application of linear and time-invariant encoding and decoding operations, coupled with a restricted number of microphones in the array (or ambisonic channels), results in inherent spatial resolution limitations. For example, the most popular microphone array configuration used for Ambisonics recording involves arranging four microphones in an open tetrahedral fashion, which may be used to capture the sound scene in what is referred to as *first-order* Ambisonics. Here, the captured sound-field is represented by an omnidirectional signal and three orthogonal *figure-of-eight* (dipole) signals. A conventional Ambisonic decoder is then tasked with linearly combining these signals to produce signals with different spatial patterns, such that, when these signals are routed to the loudspeaker array or headphones, the appropriate perceptual cues are delivered to the listener. In the ideal case, regardless of where the sound sources were positioned relative to the recording position, the reconstruction of the sound-field on the playback side should exactly match the sound-field that was recorded. However, when working with limited spatial information, it is not possible to obtain a perfect reconstruction for all directions, and thus a compromise must be made. Often, this compromise is to ensure that the squared spatial reconstruction error averaged over a dense spherical grid of directions is minimised.

A consequence of low spatial resolution, is that there is often a considerable amount of overlap between the spatial patterns of the signals delivered over the playback setup. Therefore, sounds that were initially concentrated in one particular direction during the recording, may instead be reproduced over multiple loudspeakers surrounding the true direction of the sound. Perceptually-speaking, this inherent *coherent spreading* of sound sources can lead to localisation ambiguities and to a small suitable listening area (or *sweet-spot*). Additionally, since a large portion of each loudspeaker signal may also be spread and played out of many nearby loudspeakers, with the same phase, timbral colourations may occur. On the other hand, *spatially incoherent* sounds, which should arrive at the listening point from all directions, and with random phase, such as diffuse reverberation, may erroneously become less diffuse, which could sound unnatural to the listener. All of these issues may be alleviated to some degree by increasing the number of microphones in the array, since this permits an increase in the Ambisonic order/spatial resolution and, in turn, results in a reduction in the average squared reconstruction errors. However, it should be highlighted that there are few commercially available microphone arrays capable of capturing higher-order Ambisonics, and they

are often expensive, or otherwise limited to capturing these Ambisonic components within narrow frequency bandwidths.

The perceptual issues which arise when coupling linear and time-invariant recording and reproduction methods with a limited number of microphones, has been a key motivation behind proposals to develop signal-dependent alternatives. These alternative methods are based on the philosophy that, rather than relying purely on the knowledge of the microphone array and playback setup specifications to derive fixed multi-channel filters, one may also observe the relationships between captured signals, in order to gain insights into the composition of the sound scene. This information may then help facilitate a more suitable mapping of the microphone array (or Ambisonic) signals to the playback setup, such that the aforementioned perceptual issues may be mitigated. A simple example of this involves the observation of the inter-channel relationships between first-order Ambisonic signals. If it could be established that the recording comprised a single stationary sound source in a free-field (anechoic) environment, and its direction relative to the recording point could be subsequently ascertained, then one could simply route the omnidirectional signal (of the Ambisonics representation) directly to a loudspeaker coinciding with this estimated source direction. In doing so, one would effectively obtain the maximum attainable spatial resolution for that playback setup. It is in essence based upon these kinds of principles, where assumptions regarding the composition of a sound scene may be made, and signal-dependent rendering techniques derived, in an attempt to go beyond what would otherwise be possible with a signal-independent solution.

Imposing assumptions regarding the composition of a sound scene is more formally referred to as adopting a *sound-field model*. The chosen sound-field model may then serve as a foundation, dictating which spatial parameters need to be estimated, and how they should be subsequently used to synthesise the target loudspeaker or binaural signals most optimally. It is noted that this type of processing is commonly referred to as parametric spatial audio reproduction, since it involves an attempt to describe the sound scene using a (often sparse) set of spatial parameters. These spatial parameters are also usually interpretable in isolation, such as the direction of incidence of a sound source.

The first so-called *parametric* sound-field reproduction method was the Spatial Impulse Response Rendering (SIRR) [37] method, which operates on microphone array room impulse responses encoded into the Ambisonics format. Shortly thereafter, a similar formulation which instead targeted the rendering of Ambisonic signals and known as Directional Audio Coding (DirAC) [51] was introduced. Both SIRR and DirAC operate in a similar manner to the simple example described above, except that the single source model is applied independently across time and frequency. In practice, the direction-of-arrival (DoA) of a reflection/source is estimated in

time windows every few milliseconds, and over frequencies corresponding to a perceptually meaningful scale. The omnidirectional signal is then reproduced over the playback setup using amplitude panning, which allows the signal to be spatially interpolated in cases where the estimated DoA falls in-between two or three loudspeakers or available binaural filters. Since the source signals are no longer spread to many loudspeakers surrounding the true source direction, as would otherwise be the case with a conventional lower-order Ambisonics rendering pipeline, this so-called *direct-stream* rendering of SIRR and DirAC may largely circumvent the aforementioned localisation ambiguity issues.

One additional key benefit of the SIRR and DirAC methods then lies in the authors' acknowledgement that it may not be appropriate for all sound scene components to be reproduced in the same manner. For example, diffuse reverberation, which naturally arrives with random phase from many directions surrounding the listener, may be more appropriately reproduced using an alternative rendering strategy. This is the reason why a measure of the sound-field *diffuseness* is also estimated per time-frequency index by these methods. The intention is for this diffuseness measure to be maximised in cases where the sound-field contains only ambient sounds, and minimised when it instead contains a single dominant sound source. The authors proposed that time-frequency indices found to correspond to ambient sounds may be more faithfully reproduced by: first modulating the omnidirectional signal with the diffuseness estimate, routing it to all channels in the playback setup, and then applying operations which decorrelate the signals. Finally, by modulating the direct-stream by the complement of this diffuseness estimate and summing it with the decorrelated signals (also referred to as the *diffuse-stream*), the SIRR and DirAC methods seek to largely mitigate the other perceptual issues associated with traditional lower-order signal-independent approaches; such as a loss of perceived spaciousness and envelopment.

In principle, all parametric rendering methods are capable of attaining the maximum possible spatial resolution afforded by the target playback setup provided that: 1) the input sound scene is compatible with the adopted sound-field model, and 2) the spatial analysis and synthesis techniques mitigate any potential estimation and rendering errors. In practice, however, sound-field model mismatches can occur in reality, and achieving perfectly accurate spatial estimators and rendering techniques for all scenarios is not possible, which can lead to audible artefacts being introduced into the output. Therefore, the success of a parametric method is mainly dependent upon making assumptions which are sufficiently general, so as to be largely applicable to a number of different possible sound scene scenarios. Additionally, care must be taken, in order to design and implement robust and accurate spatial analysis and synthesis techniques.

At a first glance, the imposition of a single source assumption, as is the case with SIRR and DirAC, may be viewed as an easily violated and unrealistic approximation. However, it is highlighted that for many simple and moderately complex scenes with multiple active sound sources, it may be uncommon for all sources to be active at the same time and in the same narrow frequency band. There also exists some degree of perceptual masking of rendering artefacts when many sources are active, or if there is a sufficiently high amount of reverberation in the recording. Therefore, despite its apparent limitation, the single-source model has been shown to be effective when rendering a number of diverse sound scenes in formal listening tests [49]. It is therefore highlighted that these methods usually only falter when presented with a small number of temporally and spectrally overlapping sources, while also in an acoustically dry environment.

Over the last decade and a half, research into parametric rendering methods has focused on: 1) better handing of problematic sound scenes, by adopting more general sound-field models and rendering techniques, which can also take advantage of microphone arrays containing many more microphones (or using higher-order Ambisonics) as input; 2) the deployment of different rendering strategies to overcome challenges arising from the use of sub-optimal microphone arrays, such as those integrated into mobile devices (where spatial sound recording is not necessarily their primary or only function); 3) how the parameterisation of the input sound scene, and the estimated spatial parameters, may be manipulated in an advantageous way, in order to alter the recorded sound scene in a desired manner, and to explore potential in the emerging area of so-called *superhuman* or augmented hearing technologies; and 4) investigating how so-called *six-degrees-of-freedom* rendering may be facilitated; whereby both the listener's head orientation and position are taken into account during the reproduction; thus, allowing a listener to navigate around the recording point using one microphone array, or in-between many recording points when using multiple microphone arrays.

In this thesis work, the use of a number of different parametric sound-field models and rendering techniques were explored, with contributions made to all four lines of research described above. Publication I details a higher-order formulation of SIRR, capable of leveraging the additional spatial information found in higher-order Ambisonic room impulse responses. Publication II concerns a parametric Ambisonic encoding approach, which is able to generalise to arbitrary microphone array geometries and deliver ambisonic signals over a wider frequency bandwidth, and with higher spatial resolution, than would otherwise be possible with conventional signal-independent encoders. Publication III concerns the processing of head-worn microphone array signals for superhuman and augmented hearing applications; whereby, ultrasonic sounds are rendered such that they are both audible and localisable (in the estimated direction) by a listener. Para-

metric sound-field modifications and spatial audio effects were explored in Publication IV, which may be used to augment the listener's experience of a captured sound scene; including six-degrees-of-freedom rendering using a single microphone array. Another six-degrees-of-freedom rendering system was then proposed in Publication V, which instead utilises multiple Ambisonic receivers in order to capture the sound scene from multiple perspectives; thus, obtaining additional information, which may be leveraged to overcome certain limitations associated with single microphone array based solutions.

2. Summary of Contributions

This chapter comprises a series of short serialised introductions and summaries of the publications that are appended to this thesis. For a more detailed introduction into microphone array based parametric spatial audio technologies, the reader is referred to [49, 40, 13, 64, 29, 1, 36].

2.1 Rendering higher-order Ambisonic room impulse responses

The first contribution of this thesis is related to the task of imparting the spatial characteristics of real spaces onto monophonic input signals. Here, the intention is to first capture the acoustical properties of a space of interest through an Ambisonic room impulse response (RIR) measurement, and use it to synthesise RIRs that instead correspond to a target playback setup. By then convolving the synthesised RIRs with a monophonic input signal and playing it over the target playback setup, the acoustical properties of the captured space (given the source/receiver combination used for the RIR measurement) may be imparted onto the signal and delivered to the listener. There are a handful of linear and time-invariant solutions available for the task of estimating these playback RIRs; many of which may be found within the vast literature surrounding the Ambisonics framework [19, 72, 71, 73, 15]. It is noted, however, that limitations associated with a signal-independent mapping of lower-order Ambisonic RIRs to the target playback format have been shown to lead to perceptual deficiencies; including: localisation ambiguities, poor externalisation of sound sources, timbral colouration issues (such as comb-filtering), and a diminished sense of spaciousness and envelopment [54, 10, 3, 7, 59]. These issues are all largely down to the same inherent problem, which is that the low spatial resolution results in a large degree of overlap in the spatial patterns of the signals delivered over the playback setup, which results in sound-field reconstruction errors. Here, point-sources may instead be erroneously reproduced in a spatially coherent manner over an extended region on the sphere, leading to localisation ambiguities. Additionally, diffuse sounds

may be reproduced in a non-diffuse manner, thus contributing to the other perceptual issues of concern.

As discussed in the introduction chapter, parametric signal-dependent rendering alternatives aim to address these issues by imposing assumptions regarding the composition of the RIR; estimating meaningful spatial parameters over time and frequency, and using this information to map the input RIR to the target loudspeaker setup in an adaptive and more informed manner. One early study investigating the direction-of-arrival (DoA) estimation of reflections within RIRs, which is a key spatial parameter utilised by the vast majority of parametric methods, was conducted in [67]. Here, broad-band DoAs were estimated using either the intensity vector derived from first-order Ambisonic input, or based on the time-differences of arrival using an open spherical arrangement of omnidirectional microphones. The intention of this spatial analysis was for the subsequent visualisation of the spatial RIR. However, by applying this intensity-vector based DoA estimation across both time and frequency, and using it to spatialise the omnidirectional component using the Vector-Base Amplitude Panning (VBAP) approach [47], one essentially describes the direct-stream rendering conducted by the Spatial Impulse Response Rendering (SIRR) method [37, 50]. This was the first method proposed for the task of reproducing spatial RIRs. Whereas, the alternative spatial decomposition method (SDM) [62] first applies the broad-band time-difference of arrival based DoA estimation, and uses the resulting estimates to quantise an omnidirectional signal to the nearest loudspeaker. These two methods perhaps represent the most popular spatial RIR rendering approaches, and have since received multiple extensions and optimisations [61, 12, 68, 2, 23]; including in the present contribution.

One other differentiating characteristic of the SIRR method is in its separate handling of diffuse sound components. Here, accompanying the DoA estimates, a diffuseness measure is also estimated across time and frequency, and used to modulate the amplitude of a dedicated diffuse-stream renderer. This diffuse-stream rendering is realised by replicating the omnidirectional signal for all output channels, followed by the application of decorrelation operations to ensure a spatially incoherent (i.e., a diffuse) reproduction. The amplitude of the direct-stream is then modulated by the complement of this diffuseness measure, and combined with the diffuse-stream. In the ideal case, the diffuseness estimate should approach zero if a time-frequency tile contains a single reflection or direct sound, and thus the omnidirectional component will be spatialised over the target setup only in the estimated DoA. This processing would therefore address the coherent source spreading and localisation ambiguity problems associated with lower-order signal-independent based methods. Whereas, the diffuseness estimate should be maximised for time-frequency tiles that contain only ambient sounds, and thus the spatially incoherent

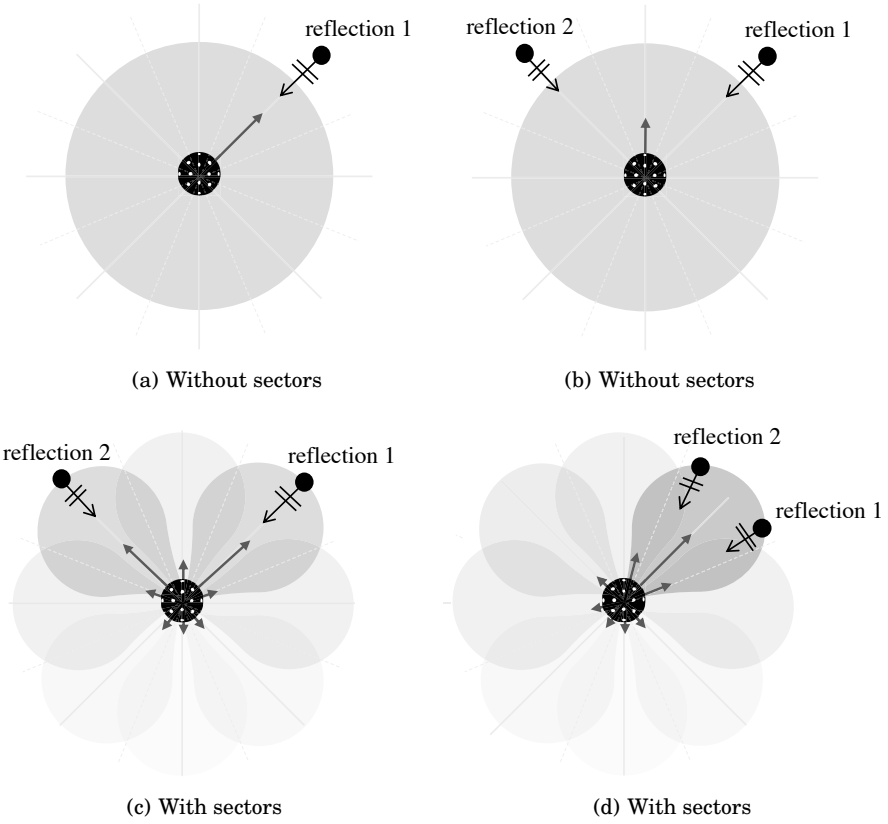

Figure 2.1. Illustrations of intensity-based DoA estimation both with and without the use of sector beamforming. The arrows protruding from the array depict the opposite direction of the flow of acoustical energy [16], which are assumed to correspond to the DoA of a reflection.

rendering approach, which is applied only to these diffuse parts of the response, should retain an appropriate perceived level of spaciousness and envelopment. The results of formal perceptual studies have indeed demonstrated that the SIRR method can produce output responses which are closer to reference responses [50], when compared to those obtained via linear Ambisonics decoding.

Previous extensions to the SIRR and SDM methods primarily looked towards adaptation to accommodate direct-to-binaural playback [68, 2, 24] (rather than binauralising loudspeaker renders), or to permit user modifications [12]. These extensions still adopt the single-source assumption per time window or per time-frequency tile, which is likely to be suitable for the early part of the response. However, over time, in those increasingly common cases where multiple reflections are present within the same time window and frequency band, and when intensity-based DoA estimation is used, the DoA estimate will be erroneously assigned in a direction some-

where in-between these reflections; thus, leading to potential localisation errors. Figure 2.1(b) illustrates this problem, given a scenario involving two reflections of equal sound energy. Addressing such issues was the primary motivation for the contribution described in Publication I. Here, a multi-source extension to the SIRR method was proposed, by adopting the sector-based spatial analysis described in [43]. This sector-based analysis has previously been used by higher-order formulations of DirAC [46, 42], and also by the present author for sound-field visualisation purposes [33]. The main philosophy behind the approach is that, by partitioning the sound-field into spatially selective *sectors*, the intensity-based DoA estimation may be conducted independently for each directionally-constrained region on the sphere. Should reflections arrive simultaneously in time and frequency, then, provided that they fall within their own sector, their local DoA estimates should be more accurate. This type of analysis is illustrated in Figure 2.1(c). It is noted, however, that should multiple reflections land within the same sector, then the same issues encountered by the first-order SIRR formulation will occur. However, the angular error of the DoA will likely be lower, as this undesired behaviour would be largely restricted to within that sector; as illustrated in Figure 2.1(d). The sector-based analysis also scales with input Ambisonic order, with higher-orders permitting greater subdivision of the sphere into narrower sectors, which can further alleviate such issues.

The proposed higher-order SIRR formulation was evaluated by conducting a formal perceptual study using a 64-channel spherical loudspeaker array in an anechoic chamber. Simulated loudspeaker RIRs served as the reference cases, which were also encoded into first-, third-, and fifth-order Ambisonics. This allowed renders using the proposed higher-order SIRR formulation to be compared against those produced using a signal-independent Ambisonic decoder. The results of the listening test showed that first-order SIRR was rated as being closer to the reference than linear first-order Ambisonics rendering in all cases, and also rated higher than third-order Ambisonics in the majority of cases. Third- and fifth-order SIRR were shown to provide incremental perceptual improvements over first-order SIRR (and over their signal-independent counterparts) for an auditorium. Third-order SIRR improvements over third-order Ambisonics were demonstrated for a concert hall case; although, all three SIRR cases were rated similarly. Fifth-order SIRR was then rated similarly or slightly higher or lower than fifth-order Ambisonics, which may indicate diminishing returns for the parametric method at higher-orders; i.e., where the spatial resolution may already be sufficiently high, that the perceptual issues of signal-independent approaches become less problematic.

2.2 Ambisonic encoding of arbitrary microphone arrays

The Ambisonics format has gained increased traction within the consumer space in recent years, which is largely due to emerging AR/VR devices and public interest in 360 degree immersive media. One of the main reasons for its popularity is that it essentially serves as a common intermediary, allowing the decoupling of the capture device(s) from the reproduction device(s); i.e., an Ambisonic decoder does not (typically) need information regarding the specific microphone array used to acquire the Ambisonic signals, and the encoder does not require knowledge concerning the possible playback system(s). The format can also facilitate well-defined and computationally efficient sound-field rotations [25, 52], in order to deliver a three-degrees-of-freedom rendering over head-tracked headphones. Furthermore, since it is based on spherical harmonics, which are orthogonal basis functions, the format represents the most efficient means of storing a 360 degree sound scene from one point in space (without favouring particular directions, or having to resort to a parameterisation of the sound-field). It may therefore be argued that the perceptual limitations described in Section 2.1 are not incurred by the use of the Ambisonics format itself, but are a product of applying signal-independent methods for both the encoding and decoding stages. While such processing paradigms indeed cannot introduce time-varying artefacts, and may always represent the most computationally efficient option, they are inherently limited in the maximum spatial resolution they can deliver over a playback system.

Within audio engineering contexts, spherical microphone arrays (SMAs) are often employed for the task of capturing Ambisonic signals. This is largely due to their ability to deliver consistent spatial resolution for all directions on the sphere, and because traditional encoding approaches [39, 27, 52, 70, 32] may also be conveniently derived based on analytical descriptions of the array geometry and its construction [63, 66]. The most popular SMA configuration is of an open tetrahedral arrangement of cardioid sensors, which is able to capture up to first-order Ambisonic signals when using these traditional linear encoding approaches. Due to physical limitations, however, the performance of this encoding is both frequency- and order-dependent, which means that the resulting omnidirectional signal and three dipole signals will only exhibit their intended patterns, and be useable (spatially-speaking), within certain frequency ranges. Above the upper limit of these ranges, the patterns will succumb to spatial aliasing, which is where the intended patterns become corrupted by more complex spatial patterns. Whereas, below the lower limit, the patterns become less spatially selective and/or contaminated by sensor noise. These latter issues also become more prevalent at higher-orders.

When subsequently reproducing lower-order Ambisonic signals over the playback setup using a signal-independent decoder, the perceptual issues

of concern described in the previous subsection can arise. Therefore, over the last two decades, many proposals for signal-dependent alternatives for the decoding part of the Ambisonics pipeline have sought to address these issues [48, 5, 65, 45, 31, 57]. However, very few studies have sought to tackle the perceptual problems through a signal-dependent encoding solution on the capture side. Those that have, focused on extending the usable frequency bandwidth beyond the spatial aliasing frequency for first-order [56] and higher-order [30, 44] capture capable SMAs. Due to the popularity of the format, however, a future need may arise for Ambisonic signals to be captured by devices whose primary function is not sound-field recording. These may include smartphones, wearable devices related to AR applications, and 360-degree video cameras. Prior to the present contribution, however, no generalised signal-dependent solution existed for the Ambisonic encoding of arbitrary microphone arrays.

It is first acknowledged that all parametric methods can experience sound-field model violations, which can potentially incur audible artefacts. In practice, such artefacts can be addressed, to some degree, through tuning the rendering parameters; such as the frequency resolution and temporal averaging coefficients. Often, this is in order to favour a more robust operation, at the expense of reducing the responsiveness of the system. However, alternatively, one can also explore the adoption of more expansive sound-field models, which can often accommodate aspects such as multiple simultaneous plane-waves per time-frequency. While such models (and associated rendering architectures) may require a more complicated implementation, they may also have the capacity to mitigate certain artefacts, which would otherwise not be possible to address through parameter tuning alone. Therefore, given the more limited microphone array configurations considered within the present contribution, an alternative sound-field model and parametric rendering architecture (compared to the SIRR and DirAC methods) was adopted.

The sound-field model employed in Publication II assumes the presence of a variable number of sound sources and an anisotropic diffuse component per time-frequency. In order to parameterise the sound scene, a detection algorithm is first used to establish how many sound sources are active [21, 11, 26]; this is followed by the estimation of the respective DoAs of the sources [55, 14]. Signal-dependent spatial filters (beamformers) are then used to isolate the source signals [18], which are then subtracted from the input microphone array recording, in order to also obtain an estimate of the anisotropic diffuse sounds captured by the array. The framework may therefore be viewed as a re-formulation of the Coding and Multi-Parameterisation of Ambisonic Sound Scenes (COMPASS) [45] method, such that it may operate directly based upon microphone array signals. The sound source signals are then encoded into the Ambisonic format by simply applying the spherical harmonic weights corresponding to the

respective DoAs. It is also important to emphasise that the order of this encoding is not constrained to a maximum Ambisonic order, which would otherwise be dictated by the number of microphones for a traditional signal-independent encoding. The microphone array domain representation of the diffuse sounds is then also converted into the Ambisonics format, by first projecting the residual recording onto a uniformly distributed spherical grid of directions, subjecting these signals to decorrelation operations, and then encoding them at the same target Ambisonic order.

To evaluate the proposed Ambisonic encoding approach, a seven-sensor microphone array affixed to a head-mounted display, which was worn by a manikin, was first simulated. Here, the capture characteristics of the array were obtained for a dense spherical grid of directions. This facilitated the creation of simulated microphone array recordings, which could be encoded into the Ambisonics format using a traditional signal-independent encoder, and also the proposed parametric encoder. Note that this particular array layout was chosen as it represents a possible future configuration within the context of AR devices, where conventional signal-independent encoding methods would be especially limited. The microphone array was encoded into first-order Ambisonics using the selected signal-independent encoder, and also encoded into fifth-order Ambisonics using the proposed method. The encoded signals were then decoded using a signal-independent approach for binaural playback [58], and compared against a reference scenario, which was obtained through the decoding of perfect (idealised) fifth-order Ambisonic signals provided by the simulator. For additional insight, the popular tetrahedral SMA configuration was also encoded into first- and fifth-order Ambisonic signals using the signal-independent and proposed parametric method, respectively. Based upon the results of a formal perceptual study, the decoded output of the proposed parametric encoding approach was shown to be closer to the reference; both when using the head-worn microphone array, and the tetrahedral SMA. Additionally, the perceived performance when using these two contrasting array configurations was shown to be similar; thus, demonstrating the general nature of the proposed encoding solution.

2.3 Superhuman spatial hearing for ultrasonic sound sources

Humans are sensitive to a wide range of frequencies. However, there is an upper limit, which is approximately 20 kHz for young people, which gradually lowers with increasing age. The frequencies above 20 kHz are commonly referred to as ultrasonic frequencies, and many animals, such as bats, rodents, insects, reptiles and amphibians, produce strong vocalisations in the ultrasonic range [53]. Man-made devices may also generate ultrasonic sounds during their normal or abnormal operation

(such as gas leaks in pipes [60]). Prior to the present contribution, there existed signal processing techniques able to bring these ultrasonic signals down to the audible range [6, 69]. For example, bats are often monitored using specific detectors [4], which can play back the down-pitch-shifted sounds through a miniature loudspeaker. However, while the sounds produced by these devices are audible to the human listener, such devices do not permit the simultaneous perception of the direction of the recorded ultrasonic sound sources.

In Publication III, a head-worn microphone array was constructed based on the use of six ultrasonic microphones. These were flush-mounted onto a rigid spherical baffle, which was approximately 11 mm in diameter. The array was then suspended in front of the head of the listener by mounting it on the end of a short rod, which was affixed to the headband of a pair of headphones worn by the listener. The intention of this study was to then assess the localisation abilities of listeners, when ultrasonic sound sources were captured by the array, and their signals down pitch-shifted and spatialised over the same pair of headphones in real-time. In the present study, the frequency range 20-96 kHz of one microphone array signal was down pitch-shifted to 2.5-12 kHz (i.e., down three octaves) and then spatialised based on a broad-band DoA estimate. This broad-band DoA estimate was acquired by averaging narrow-band DoA estimates [41] made over the frequency range: 20-55 kHz. Since the listener would otherwise be unable to perceive or localise the ultrasonic sound sources, the developed system may be described as bestowing *superhuman* spatial hearing abilities onto the wearer. The results of the listening test indicate that the subjects were, on average, able to perceive and localise ultrasonic sound sources within 4 degrees on the horizontal plane. As similar processing is adopted within hearing aid devices [20], this contribution represented the first study to exaggerate this concept, in order to extend the spatial hearing capabilities of normal hearing subjects.

2.4 Spatial audio effects and sound-field modifications

One other important aspect of spatial audio technologies, which may also have implications within future augmented hearing devices, is the optional ability to manipulate the spatial properties of the captured sound scene prior to its reproduction. In Publication IV, the aforementioned COMPASS [45] sound-field model and rendering framework, which operates based upon Ambisonics signals as input, was explored for realising a number of spatial audio effects and sound-field modifications. Here, the intention was to offer the audio engineer a degree of control over the manipulation of the sound scene that would otherwise not be possible, when existing signal-independent sound-field modification approaches are employed. As

described in Section 2.2, the framework operates by decomposing the input sound scene into individual sound source components, which are accompanied by an anisotropic residual component encapsulating ambient sounds. Since this latter component is also represented in the Ambisonics format, it is highlighted that any existing signal-independent spatial audio effect or sound-field modification intended for Ambisonic input [28], may also be applied to this residual component; i.e., with that effect or modification only affecting sounds which the framework deemed to be ambient. In this contribution, however, specific focus was placed on the effects and modifications that can be realised through manipulation of the estimated spatial parameters, prior to rendering the modified scene for the target playback setup.

Due to the decoupling of the direct sounds from the anisotropic ambient sounds, one intuitive spatial audio effect involves biasing the balance of these two audio streams. This may be performed to reduce the level of spatially ambiguous or extended background sources in the scene, which have mostly diffuse characteristics. Alternatively, this captured and isolated ambience of the scene may be exaggerated for creative purposes. Other spatial audio effects include those based on applying directional transformations onto the source directions (prior to spatialising their signals), and/or the application of direction-dependent gain factors. These transformations and effects are illustrated in Figure 2.2.

The first illustrated transformation is sound-field rotation, whereby source beamformers remain steered towards the estimated DoAs, but the spatialisation gains applied to their signals instead correspond to rotated DoAs. Although sound-field rotations may, as an alternative, be realised through the application of existing, well-defined rotation matrices applied directly onto the Ambisonic signals [25]; it is highlighted that the ability to apply rotations based upon the parameterised and decomposed source signals (i.e. after the scene analysis stage) has some practical benefits. Namely, by carrying out the spatial analysis only once (e.g., on a remote server), and subsequently being able to cater for multiple listeners with different head orientations on different device(s), this alternative means of accommodating for sound-field rotations represents an efficient solution, if such a parametric framework were to be adopted at scale.

The second directional transformation explored was the *warping* of directional sounds. Here, the intention is for an audio engineer to first place markers on the unit-sphere, with the instantaneous DoA estimates subsequently being directionally biased towards these markers, depending upon their proximity to them; i.e., if the DoA is near to a marker, then it may be more drastically *pulled* towards it, whereas, if the DoA is far from a marker, then it may remain fixed or move only slightly. Other than creative uses, if the markers correspond to the known source directions (obtained, for example, through other modalities, such as by analysing

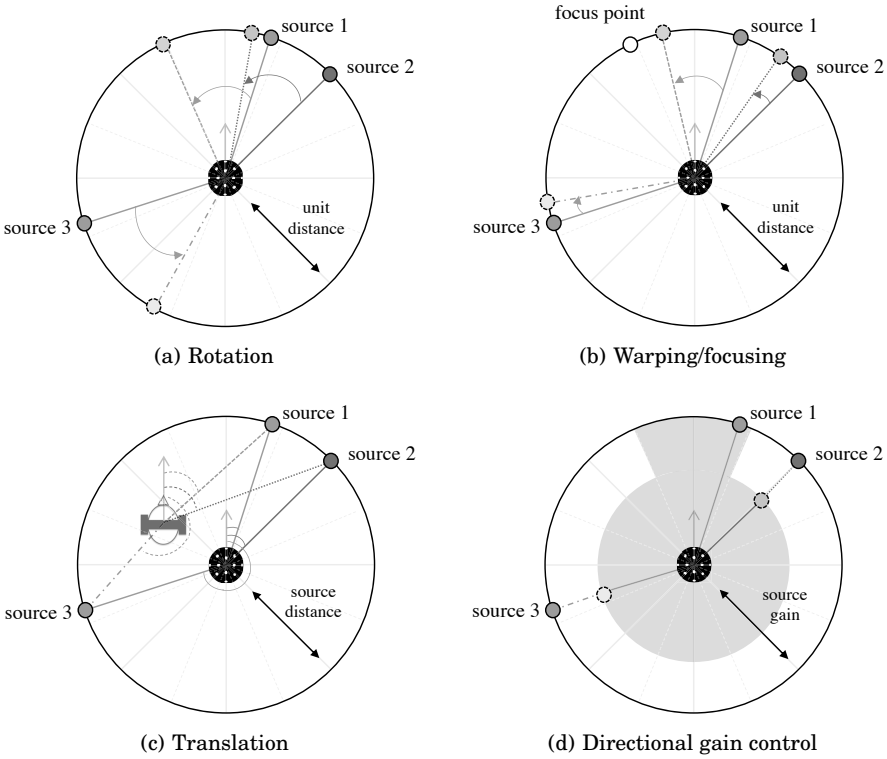

Figure 2.2. Illustrations of how the sound scene may be manipulated based on simple parameter adjustments, either realised through directional transformations of the DoAs (a-c), or through direction-dependent gain factors (d).

simultaneous video recordings or performing optical tracking of predominant sources), then this type of transformation may also be viewed as a focusing operation; possibly alleviating potential issues of inaccurate DoA estimation, which may affect the perceived stability of the rendering.

The third effect investigated was sound-field translation, whereby the DoA estimates are projected onto a sphere with a radius corresponding to the assumed/known source distances. Through trigonometric operations, the spatialisation directions may be manipulated to account for a translated listener; thus, permitting six-degrees-of-freedom rendering using only a single microphone array recording. Source- and distance-dependent gains may also be included to further improve the perceived effect. The main limitation of this translation approach, however, is that the source distances need to be known or assumed. While distance estimation algorithms based upon microphone array recordings are available [22], their accuracy is typically limited to a few tens of centimetres from the array. Therefore, future single-point six-degrees-of-freedom approaches may need to look for alternative source distance estimators using other modalities.

The fourth illustrated effect, direction-dependent gains, is also easily realisable using the described framework, which allows a user to amplify or attenuate certain regions on the sphere. This effect was also further explored by the present author in [17], except within the context of rendering head-worn microphone array recordings.

One other effect, explored in this contribution, was inspired by traditional single-channel workflows, whereby, one audio signal is analysed and used to manipulate another audio signal. An example of such a workflow is side-chain dynamic range compression, whereby the compression of an audio signal is dictated by the analysis of a different audio signal. In this contribution, the results of a spatial analysis operation conducted upon one input sound scene are used to dictate how a second sound scene is decomposed and re-synthesised.

The remaining sound-field modifications explored in the present contribution were realised by first clustering the DoAs across frequency, and then tracking them as broad-band sound objects on the unit sphere. The tracking algorithm employed for this task was proposed by the doctoral candidate and is described in [34]. Once the directions of the active sound source(s) are known, this facilitates the application of flexible gain re-balancing, source-specific single-channel effects, and directional re-assignments of the source objects in the scene.

This contribution featured no formal evaluation, but introduced five real-time audio plugins, which serve as examples of how the described spatial audio effects may be realised in practice. One audio plugin consisted of a binaural decoder, which supports efficient reproduction for multiple simultaneous listeners; i.e., with the spatial analysis conducted once, and the synthesis techniques applied independently per listener, based on their own head-tracking data and/or individualised binaural filters. Another plugin sought to demonstrate the directional focusing effect by allowing the user to place markers on an equirectangular representation of the sphere, and then control the degree to which the DoA estimates should be *pulled* towards these markers; thus, manipulating the sound-field prior to its reproduction. A flexible sound-field editor, which also relied on the use of markers, was additionally developed. In this case, these markers may either be placed in user defined directions, or placed automatically based on the output of the adopted tracking solution. Single-channel effects and gain factors may then be (optionally) applied to the signals of the individually tracked sources, before the sound scene is reconstructed and delivered in the Ambisonics format.

2.5 Six-degrees-of-freedom rendering of spatial sound scenes

In Publication V, a complete rendering system, which supports six-degrees-of-freedom listener movement, was proposed and evaluated. Contrary to the six-degrees-of-freedom rendering described in the previous subsection, which was based upon a single microphone array as input, the system developed in this contribution employs the use of multiple microphone arrays. These microphone arrays may be distributed arbitrarily over an extended spatial area of interest. By observing the sound scene from multiple perspectives, the system is able to leverage this additional information to overcome some of the issues with the single microphone array approach. Namely, the position of sources may be ascertained by triangulating the DoAs, as opposed to requiring the source distances to be specified to the rendering system.

The proposed system takes an object-based rendering approach, which means that prominent sound sources in the scene are identified and tracked over time, with their signals subsequently isolated by steering a broadband beamformer towards them using the nearest Ambisonic receiver. Note that this tracking was conducted in a similar manner to that described in the previous contribution, with the exception that the tracking framework was provided with Cartesian coordinates of source position estimates, as opposed to unit-length Cartesian vectors describing source directions [34]. The isolated broad-band sound object signals are then reproduced over the target playback setup, taking into account the listener orientation and position relative to the tracked source positions. For simplicity, the directivity of the sources was not estimated, nor considered during rendering; i.e., the sources were assumed to be omnidirectional. The ambient stream was then realised by subtracting the sound sources from receivers close to the listener position, in the same manner as described in [45] except with the additional application of interpolation weights. Therefore, in essence, the proposed system may be viewed as a natural multi-microphone array extension to the COMPASS method.

The proposed six-degrees-of-freedom approach was evaluated through perceptual listening tests. Here, the dataset released in [35] was employed; which comprises a number of spatial RIR measurements of a variable acoustics room located at Aalto University, Finland, with the dataset featuring seven different receiver positions and three source positions. The receivers were of a commercially available 32-channel SMA [38], and the sources were loudspeakers. The seven microphone array RIRs were then encoded into second-order Ambisonic signals and subsequently convolved with different stimuli for each source position, in order to obtain a synthetic recording of the sound scene from seven different perspectives. These seven second-order recordings were then binaurally rendered by the proposed system, and delivered over acoustically transparent headphones worn by

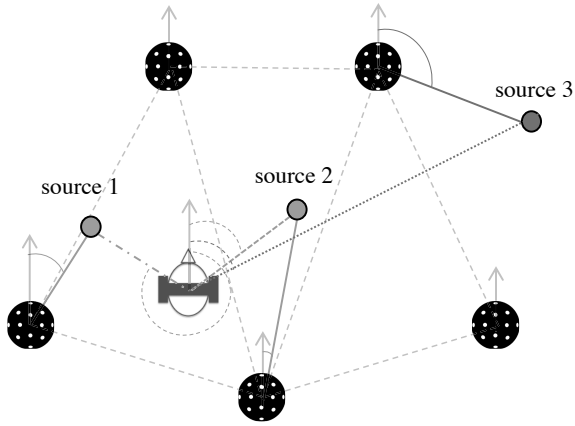

Figure 2.3. Illustration of listener translation using a distributed arrangement of multiple receivers.

a head-tracked listener located in the same variable acoustics room. The same three loudspeakers used for the original RIR measurements were also placed in those same positions; in order to allow the binaural renderings to be directly compared to the reference loudspeaker reproduction. The three test cases were of a linear baseline method, and the proposed parametric method using either the built-in tracker, or operated based upon the known source positions. The listening test results showed that, in the vast majority of instances, both parametric test cases were rated higher than their signal-independent counterpart. Additionally, the two parametric cases were also rated similarly, suggesting that the source tracking solution was sufficiently robust for the sound scenes used for the perceptual study.

References

[1] Jukka Ahonen. *Microphone front-ends for spatial sound analysis and synthesis with Directional Audio Coding*. Doctoral Dissertation, Aalto University, 2013.

[2] Sebastià V Amengual Garí, Johannes M Arend, Paul T Calamia, and Philip W Robinson. Optimizations of the spatial decomposition method for binaural reproduction. *Journal of the Audio Engineering Society*, 68(12):959–976, 2021.

[3] Amir Avni, Jens Ahrens, Matthias Geier, Sascha Spors, Hagen Wierstorf, and Boaz Rafaely. Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution. *The Journal of the Acoustical Society of America*, 133(5):2711–2721, 2013.

[4] Michel Barataud. Acoustic ecology of european bats. *Species identification and studies of their habitats and foraging behaviour. Biotope Editions, Mèze*, 2015.

[5] Svein Berge and Natasha Barrett. High angular resolution planewave expansion. In *Proc. of the 2nd Int. Symp. on Ambisonics and Spherical Acoustics*, pages 6–7, 2010.

[6] Stephan M Bernsee. Pitch shifting using the Fourier transform. *The DSP Dimension, http://blogs. zynaptiz. com/bernsee/pitch-shifting-using-the-ft*, 1999.

[7] Stéphanie Bertet, Jérôme Daniel, Etienne Parizet, and Olivier Warusfel. Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources. *Acta Acustica united with Acustica*, 99(4):642–657, 2013.

[8] Jens Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[9] Alan D Blumlein. British patent specification 394,325 (improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems). *Journal of the Audio Engineering Society*, 6(2):91–130, 1958.

[10] Sebastian Braun and Matthias Frank. Localization of 3D ambisonic recordings and ambisonic virtual sources. In *1st International Conference on Spatial Audio, (Detmold)*, 2011.

[11] Weiguo Chen, Kon Max Wong, and James P Reilly. Detection of the number of signals: A predicted eigen-threshold approach. *IEEE Trans. Signal Processing*, 39(5):1088–1098, 1991.

[12] Philip Coleman, Andreas Franck, P Jackson, R Hughes, Luca Remaggi, and Frank Melchior. Object-based reverberation for spatial audio. *Journal of the Audio Engineering Society*, 65(1/2):66–77, 2017.

[13] Symeon Delikaris-Manias. *Parametric spatial audio processing utilising compact microphone arrays*. Doctoral Dissertation, Aalto University, 2017.

[14] Joseph H DiBiase, Harvey F Silverman, and Michael S Brandstein. Robust localization in reverberant rooms. In *Microphone arrays*, pages 157–180. Springer, 2001.

[15] N Epain, CT Jin, and F Zotter. Ambisonic decoding with constant angular spread. *Acta Acustica united with Acustica*, 100(5):928–936, 2014.

[16] Frank J Fahy and Vincent Salmon. Sound intensity. *The Journal of the Acoustical Society of America*, 88(4):2044–2045, 1990.

[17] Janani Fernandez, Leo McCormack, Petteri Hyvärinen, Archontis Politis, and Ville Pulkki. A spatial enhancement approach for binaural rendering of head-worn microphone arrays. In *24th International Congress on Acoustics (ICA)*, 2022.

[18] Otis Lamont Frost. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8):926–935, 1972.

[19] Michael A Gerzon. Periphony: With-height sound reproduction. *Journal of the audio engineering society*, 21(1):2–10, 1973.

[20] V Hamacher, J Chalupper, Joachim Eggers, Eghart Fischer, Ulrich Kornagel, Henning Puder, and U Rass. Signal processing in high-end hearing aids: State of the art, challenges, and future trends. *EURASIP Journal on Advances in Signal Processing*, 2005(18):1–15, 2005.

[21] Keyong Han and Arye Nehorai. Improved source number detection and direction estimation with nested arrays and ULAs using jackknifing. *IEEE Trans. Signal Processing*, 61(23):6118–6128, 2013.

[22] Adrian Herzog and Emanuël AP Habets. Distance estimation in the spherical harmonic domain using the spherical wave model. *Applied Acoustics*, 193:108733, 2022.

[23] Elias Hoffbauer and Matthias Frank. Four-directional ambisonic spatial decomposition method with reduced temporal artifacts. *Journal of the Audio Engineering Society*, 70(12):1002–1014, 2022.

[24] Christoph Hold, Leo McCormack, and Ville Pulkki. Parametric binaural reproduction of higher-order spatial impulse responses. In *24th International Congress on Acoustics (ICA)*, 2022.

[25] Joseph Ivanic and Klaus Ruedenberg. Rotation matrices for real spherical harmonics. direct determination by recursion. *The Journal of Physical Chemistry A*, 102(45):9099–9100, 1998.

[26] Jong-Shiann Jiang and Mary Ann Ingram. Robust detection of number of sources using the transformed rotational matrix. In *2004 IEEE Wireless Communications and Networking Conference (IEEE Cat. No. 04TH8733)*, volume 1, pages 501–506. IEEE, 2004.

[27] Craig T Jin, Nicolas Epain, and Abhaya Parthy. Design, optimization and evaluation of a dual-radius spherical microphone array. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 22(1):193–204, 2014.

[28] Matthias Kronlachner and Franz Zotter. Spatial transformations for the enhancement of ambisonic recordings. In *Proceedings of the 2nd International Conference on Spatial Audio, Erlangen*, 2014.

[29] Mikko-Ville Laitinen. *Techniques for versatile spatial-audio reproduction in time-frequency domain*. Doctoral Dissertation, Aalto University, 2014.

[30] Jing Lin, Xihong Wu, and Tianshu Qu. Anti spatial aliasing HOA encoding method based on aliasing projection matrix. In *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*, pages 321–325. IEEE, 2020.

[31] Leo McCormack and Symeon Delikaris-Manias. Parametric first-order ambisonic decoding for headphones utilising the cross-pattern coherence algorithm. In *EAA Spatial Audio Signal Processing Symposium*, pages 173–178, 2019.

[32] Leo McCormack, Symeon Delikaris-Manias, Angelo Farina, Daniel Pinardi, and Ville Pulkki. Real-time conversion of sensor array signals into spherical harmonic signals with applications to spatially localized sub-band sound-field analysis. In *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[33] Leo McCormack, Symeon Delikaris-Manias, Archontis Politis, Despoina Pavlidi, Angelo Farina, Daniel Pinardi, and Ville Pulkki. Applications of spatially localized active-intensity vectors for sound-field visualization. *J. Audio Engineering Society*, 67(11):840–854, 2019.

[34] Leo McCormack, Archontis Politis, Simo Särkkä, and Ville Pulkki. Real-time tracking of multiple acoustical sources utilising rao-blackwellised particle filtering. In *29th European Signal Processing Conference (EUSIPCO)*, pages 206–210. EURASIP, 2021.

[35] Thomas McKenzie, Leo McCormack, and Christoph Hold. Dataset of spatial room impulse responses in a variable acoustics room for six degrees-of-freedom rendering and analysis. *arXiv:2111.11882 [eess.AS]*, 2021.

[36] Juha Merimaa. *Analysis, synthesis, and perception of spatial sound: binaural localization modeling and multichannel loudspeaker reproduction*. Doctoral Dissertation, Helsinki University of Technology, 2006.

[37] Juha Merimaa and Ville Pulkki. Spatial impulse response rendering i: Analysis and synthesis. *Journal of the Audio Engineering Society*, 53(12):1115–1127, 2005.

[38] Jens Meyer and Gary Elko. A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–1781. IEEE, 2002.

[39] 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 *120th Convention of the AES*, pages 20–23, 2006.

[40] Archontis Politis. *Microphone array processing for parametric spatial audio techniques*. Doctoral Dissertation, Aalto University, 2016.

[41] Archontis Politis, Symeon Delikaris-Manias, and Ville Pulkki. Direction-of-arrival and diffuseness estimation above spatial aliasing for symmetrical directional microphone arrays. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 6–10. IEEE, 2015.

[42] Archontis Politis, Leo McCormack, and Ville Pulkki. Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 379–383. IEEE, 2017.

[43] Archontis Politis and Ville Pulkki. Acoustic intensity, energy-density and diffuseness estimation in a directionally-constrained region. *arXiv preprint arXiv:1609.03409*, 2016.

[44] Archontis Politis, Sakari Tervo, Tapio Lokki, and Ville Pulkki. Parametric multidirectional decomposition of microphone recordings for broadband high-order ambisonic encoding. In *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[45] Archontis Politis, Sakari Tervo, and Ville Pulkki. COMPASS: Coding and multidirectional parameterization of ambisonic sound scenes. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6802–6806, 2018.

[46] Archontis Politis, Juha Vilkamo, and Ville Pulkki. Sector-based parametric sound field reproduction in the spherical harmonic domain. *IEEE Journal of Selected Topics in Signal Processing*, 9(5):852–866, 2015.

[47] Ville Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466, 1997.

[48] Ville Pulkki. Spatial sound reproduction with directional audio coding. *J. Audio Eng. Soc.*, 55(6):503–516, 2007.

[49] Ville Pulkki, Symeon Delikaris-Manias, and Archontis Politis. *Parametric Time-Frequency Domain Spatial Audio*. John Wiley & Sons, 2017.

[50] 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.

[51] Ville Pulkki, Archontis Politis, Mikko-Ville Laitinen, Juha Vilkamo, and Jukka Ahonen. First-order directional audio coding (DirAC). In Ville Pulkki, Symeon Delikaris-Manias, and Archontis Politis, editors, *Parametric Time-Frequency Domain Spatial Audio*, pages 89–138. John Wiley & Sons, 2017.

[52] Boaz Rafaely. *Fundamentals of spherical array processing*, volume 8. Springer, 2015.

[53] Gillian Sales. *Ultrasonic communication by animals*. Springer Science & Business Media, 2012.

[54] Olli Santala, Heikki Vertanen, Jussi Pekonen, Jan Oksanen, and Ville Pulkki. Effect of listening room on audio quality in ambisonics reproduction. In *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.

[55] Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, 34(3):276–280, 1986.

[56] Christian Schörkhuber and Robert Höldrich. Signal-dependent encoding for first-order ambisonic microphones. *Fortschritte der Akustik, DAGA, Kiel*, pages 1037–1040, 2017.

[57] Christian Schörkhuber and Robert Höldrich. Linearly and quadratically constrained least-squares decoder for signal-dependent binaural rendering of ambisonic signals. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[58] Christian Schörkhuber, Markus Zaunschirm, and Robert Höldrich. Binaural rendering of Ambisonic signals via magnitude least squares. In *Proc. DAGA*, volume 44, pages 339–342, 2018.

[59] 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.

[60] Wang Tao, Wang Dongying, Pei Yu, and Fan Wei. Gas leak localization and detection method based on a multi-point ultrasonic sensor array with TDOA algorithm. *Measurement Science and Technology*, 26(9):095002, 2015.

[61] Sakari Tervo, Jukka Pätynen, Neofytos Kaplanis, Morten Lydolf, Søren Bech, and Tapio Lokki. Spatial analysis and synthesis of car audio system and car cabin acoustics with a compact microphone array. *Journal of the Audio Engineering Society*, 63(11):914–925, 2015.

[62] Sakari Tervo, Jukka Pätynen, Antti Kuusinen, and Tapio Lokki. Spatial decomposition method for room impulse responses. *Journal of the Audio Engineering Society*, 61(1/2):17–28, 2013.

[63] Heinz Teutsch. *Modal array signal processing: principles and applications of acoustic wavefield decomposition*, volume 348. Springer, 2007.

[64] Juha Vilkamo. *Perceptually motivated time-frequency processing of spatial audio*. Doctoral Dissertation, Aalto University, 2014.

[65] Andrew Wabnitz, Nicolas Epain, Alistair McEwan, and Craig Jin. Upscaling ambisonic sound scenes using compressed sensing techniques. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–4. IEEE, 2011.

[66] Earl G Williams. *Fourier acoustics: sound radiation and nearfield acoustical holography*. Academic press, 1999.

[67] 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.

[68] Markus Zaunschirm, Matthias Frank, and Franz Zotter. BRIR synthesis using first-order microphone arrays. In *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[69] Udo Zölzer. *DAFX: digital audio effects*. John Wiley & Sons, 2011.

[70] Franz Zotter. A linear-phase filter-bank approach to process rigid spherical microphone array recordings. In *Proc. IcETRAN*, Palic, Serbia, 2018.

[71] Franz Zotter and Matthias Frank. All-round ambisonic panning and decoding. *Journal of the Audio Engineering Society*, 60(10):807–820, 2012.

[72] Franz Zotter and Matthias Frank. *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer Nature, 2019.

[73] Franz Zotter, Hannes Pomberger, and Markus Noisternig. Energy-preserving ambisonic decoding. *Acta Acustica united with Acustica*, 98(1):37–47, 2012.

Aalto-DT 60/2023

ISBN 978-952-64-1244-3 (printed)
ISBN 978-952-64-1245-0 (pdf)
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 THESES