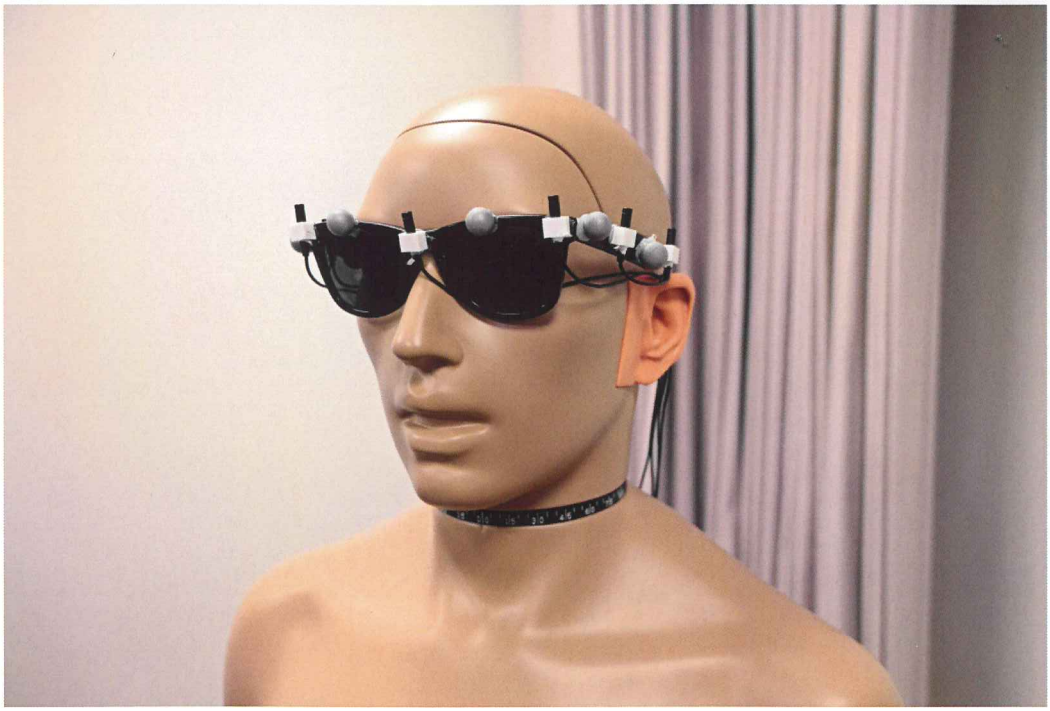
PHD THESIS



Spatial Room Impulse Response Processing for Virtual Acoustics

THOMAS DEPPISCH

DEPARTMENT OF ARCHITECTURE AND CIVIL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Spatial Room Impulse Response Processing for Virtual Acoustics

THOMAS DEPPISCH

Department of Architecture and Civil Engineering

Division of Applied Acoustics

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2026

Spatial Room Impulse Response Processing for Virtual Acoustics
THOMAS DEPPISCH
ISBN 978-91-8103-406-6
DOI 10.63959/chalmers.dt/5863

Doktorsavhandlingar vid Chalmers tekniska högskola,
Ny serie nr. 5863
ISSN 0346-718X

Department of Architecture and Civil Engineering
Division of Applied Acoustics
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone: +46 (0)31-772 1000

Acknowledgements, dedications, and similar personal statements in this thesis reflect the author's own views.

Cover:
A head-and-torso simulator wears sunglasses equipped with microphones and tracking markers.

Chalmers Digital Printing
Gothenburg, Sweden 2026

Abstract

Augmented reality (AR) and telepresence systems aim to enhance the real world with virtual elements that blend convincingly into the surrounding space. Creating virtual sound sources in this context requires presenting perceptually valid head-related and room-acoustic cues to the listener to enable a realistic spatial impression and a coherent match between the virtual acoustics and those of the physical environment. In practical AR systems, the acoustic characteristics of the environment must be estimated from available sensor signals and the virtual source rendered through acoustically transparent headphones to preserve natural sounds in the physical environment. This thesis addresses both stages of this virtual acoustic processing chain: estimation and rendering. Central to both are spatial room impulse responses (SRIRs), which describe the linear, time-invariant, and directional properties of the acoustic transfer path between a source and a receiver in an environment.

The thesis first introduces a general microphone array signal model that separates room- and array-dependent contributions using spherical or circular harmonic representations. Building on this model, a blind SRIR estimation framework is proposed that reformulates blind multichannel system identification as an informed problem through the estimation of a pseudo-reference signal. Motivated by practical AR systems that often rely on wearable devices such as head-mounted displays or smartglasses, the thesis then specifically considers microphone arrays in motion.

The second part of the thesis focuses on the binaural rendering of estimated SRIRs for headphone reproduction. An array-aware end-to-end magnitude least-squares renderer is proposed to mitigate spatio-spectral coloration caused by limited spatial sampling and regularization. As an alternative to direct rendering, the thesis investigates the separation of direct sound and early reflections from an SRIR, a common processing step in parametric SRIR-based rendering that can facilitate virtual acoustic reproduction with increased directional sharpness. Two approaches are compared: one based on a physical array signal model and another based on subspace decomposition.

Together, these contributions advance practical SRIR estimation and rendering for virtual acoustics and provide foundations for robust, wearable, and perceptually convincing augmented and virtual reality audio systems.

Keywords: Binaural Rendering, Room Impulse Response Estimation, Microphone Array, Room Acoustics, Spatial Room Impulse Response, Virtual Acoustics

Preface

I started the research that ultimately led to this thesis in the middle of the COVID-19 pandemic. While this was hardly the best time to arrive in a country I had never been to, I was incredibly lucky to have joined a lab where I felt welcome and valued while having the freedom to find my own way. Thank you, Wolfgang, for creating such an environment and thanks to all my colleagues at Applied Acoustics for bringing it to life. Thank you, Jens, for your guidance, and for being encouraging, optimistic, available, and 100% reliable. Thank you, Elin, Hannes, Jannik, and Carl, for being not just colleagues but friends. And especially thank you, Leon, for being there with me every step of the way, for countless discussions, travels, band rehearsals, and climbing sessions.

I am grateful to Meta's Reality Labs Research for funding my research. My own small project team was an absolute dream to work with. I had the unique opportunity to do academic research *and* gain insights and inspiration from my colleagues at Reality Labs, who offered a world-leading perspective on which spatial-audio problems truly matter in practice. Thank you, Sebastià, Paul, and Jens, for our many discussions and for trusting that my own ideas were worth pursuing. I appreciated that immensely.

The internships at Reality Labs were two of the most exciting and stimulating experiences during my PhD. Thanks, Sebastià and Zamir, for welcoming me and allowing me this peek behind the curtain. Thank you also to the other colleagues and co-interns at Reality Labs who made these visits special and helped me bring my latte art skills to the next level.

My scientific journey started in Graz in 2014, and this thesis would never have happened without the excellent environment and friends that I found there. Thank you to the staff at IEM who showed me how much fun research can be, and especially thank you, Franz, for sharing your endless enthusiasm and curiosity with me. Your guidance was invaluable.

I am deeply grateful to my family for their unconditional support over all these years, which enabled me to pursue my interests and play this academic game for so long. Vielen Dank! Thanks to my friends back home for still being part of my life, even though I wasn't often around. It means a lot. Thank you also to all the new friends I made. Thank you to the climbing gang. Thank you for the music, Staffan.

Last, but most importantly, thank you, Angelica, for loving and supporting me so deeply these past three years, for making my life richer and brighter. Grazie due mille.

Tommi

Göteborg, April 2026

Publications

This thesis consists of an extended summary and the following appended papers.

Paper A	T. Deppisch, J. Ahrens, S. V. Amengual Garí and P. Calamia, "Blind Estimation of Spatial Room Impulse Responses Using a Pseudo Reference Signal," *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 470–474, Seoul, Republic of Korea, 2024.
Paper B	T. Deppisch, N. Meyer-Kahlen and S. V. Amengual Garí, "Blind Identification of Binaural Room Impulse Responses From Smart Glasses," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4052–4065, 2024.
Paper C	T. Deppisch, S. V. Amengual Garí, P. Calamia and J. Ahrens, "Spatial Room Impulse Response Identification from Rotating Equatorial Microphone Arrays," *32nd European Signal Processing Conference (EUSIPCO)*, pp. 116–120, Lyon, France, 2024.
Paper D	T. Deppisch, S. V. Amengual Garí, P. Calamia and J. Ahrens, "Spatial Room Impulse Response Estimation From a Moving Microphone Array," *33rd European Signal Processing Conference (EUSIPCO)*, pp. 91–95, Palermo, Italy, 2025.
Paper E	T. Deppisch, S. V. Amengual Garí, P. Calamia and J. Ahrens, "Identification and Matching of Room Acoustics With Moving Head-Worn Microphone Arrays," *submitted to Journal on Audio, Speech, and Music Processing*, 2026.
Paper F	T. Deppisch, H. Helmholz and J. Ahrens, "End-to-End Magnitude Least Squares Binaural Rendering of Spherical Microphone Array Signals," *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–7, Remote, 2021.
Paper G	T. Deppisch, J. Ahrens, S. V. Amengual Garí and P. Calamia, "Spatial Subtraction of Reflections from Room Impulse Responses Measured with a Spherical Microphone Array," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 346–350, Remote, 2021.
Paper H	T. Deppisch, S. V. Amengual Garí, P. Calamia and J. Ahrens, "Direct and Residual Subspace Decomposition of Spatial Room Impulse Responses," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 927–942, vol. 31, 2023.

Author Contributions For all papers except PAPER G, TD had the initial idea, developed the concept, derived the methods, performed measurements, experiments, and analysis, and wrote the paper. For PAPER B, NMK performed and described the statistical analysis of the listening experiment. For PAPER G, the concept was developed together with JA. All co-authors enhanced the papers through regular discussions and feedback on the initial manuscript.

Contents

List of Symbols

Non-bold symbols denote scalar quantities or functions, bold lowercase symbols vectors, and bold uppercase symbols matrices.

a	Directional Plane Wave Density
a_n^m	SH Coefficient of the Plane Wave Density
$\boldsymbol{a}$	SH Coefficients of the Plane Wave Density up to Order N
b_n	Radial Term
$\boldsymbol{b}$	Radial Terms up to Order N
C_m	Circular Harmonics (CHs)
$\boldsymbol{d}$	Directional Responses of M Microphones
$\boldsymbol{D}$	Spatial Array Response Matrix up to Order N
$\boldsymbol{E}$	Ambisonic Encoder
f	Frequency
$\boldsymbol{h}$	Set of Head-Related Transfer Functions (HRTFs)
k	Wavenumber
l	Microphone Index
m	SH or CH Degree
M	Number of Microphones
n	SH or CH Order
$\boldsymbol{n}$	Noise
N	Maximum SH or CH Order
p	Sound Pressure
$\boldsymbol{p}_\mathrm{A}$	Anechoic Array Transfer Function
$\boldsymbol{P}_\mathrm{A}$	Set of Anechoic Array Transfer Functions
Q	Number of Directions
r	Radial Distance
$\boldsymbol{r}$	Cartesian Position
$\boldsymbol{R}$	Rotation Matrix
$\boldsymbol{R}_\mathrm{c}$	Spatial Covariance Matrix
s	Source Signal
$\hat{s}$	Pseudo-Reference Signal
$\boldsymbol{T}$	Translation Matrix

$\boldsymbol{v}$	Acoustic Array Transfer Functions
$\boldsymbol{w}_{\text{bin}}$	Binaural Rendering Filter
$\boldsymbol{W}$	Diagonal Matrix of Quadrature Weights
$\boldsymbol{x}$	Microphone Array Signals
Y_n^m	Spherical Harmonics (SHs)
$\boldsymbol{Y}$	SHs up to Order N Evaluated in a Set of Directions
ϕ	Azimuth Angle
θ	Zenith (Colatitude) Angle
Ω	A Direction Consisting of Azimuth and Zenith Angle
$\mathbb{S}_2$	Unit Sphere
$(\cdot)^*$	Complex Conjugate Operator
$(\cdot)^\top$	Transpose Operator
$(\cdot)^{\text{H}}$	Hermitian Transpose Operator
$\text{diag}\{\cdot\}$	Vector-to-Diagonal-Matrix Operator

List of Abbreviations

6DoF six degrees of freedom

ACE acoustic characterization of environments

AR augmented reality

ASM Ambisonic signal matching

ATF array transfer function

BFBR beamforming-based binaural reproduction

BRIR binaural room impulse response

BSM binaural signal matching

CH circular harmonic

DoA direction of arrival

DRR direct-to-reverberant energy ratio

eMagLS end-to-end magnitude least squares

FDN feedback delay network

GSV generalized singular value

GSVD generalized singular value decomposition

GWPE generalized weighted prediction error

HATS head and torso simulator

HRIR head-related impulse response

HRTF head-related transfer function

JND just-noticeable difference

LS least squares

magLS magnitude least squares

max-DI maximum directivity index

MPDR minimum-power distortionless response

MUSIC multiple signal classification

MVDR minimum-variance distortionless response

MWF multichannel Wiener filter

RIR room impulse response

RLS recursive least squares

RT reverberation time

RTF relative transfer function

SDM spatial decomposition method

SH spherical harmonic

SIRR spatial impulse response rendering

SMA spherical microphone array

SNR signal-to-noise ratio

SRIR spatial room impulse response

STFT short-time Fourier transform

VL virtual loudspeaker

VR virtual reality

WDRTF wearable-device related transfer function

Chapter 1

Introduction

— What does a room sound like?

Although many people can come up with an intuitive answer to this question, giving a precise answer is not straightforward. After hearing a sound in a room, we might describe it as *big*, *boomy*, *echoey*, or *dead*. We may also reach for more technical terms like *reverberant*, *dry*, *diffuse*, or *sharp*. Such descriptions capture how a space sounds to us as listeners, and they are meaningful because they combine the acoustic properties of the room with a perceived impression. They are, however, subjective and not precise enough to guide the acoustic design of an environment: two people may describe the same room differently, or differently sounding rooms with the same words.

To move from impression to repeatable measurements, acousticians ask a more concrete question: *how does the room respond to sound*? In practice, this is done by exciting the room with a standardized sound that contains all audible frequencies and recording the response of the room. From this recording, a room impulse response (RIR) is obtained. Under linear, time-invariant conditions, the RIR contains all the information needed to describe the room's acoustic behavior over time and frequency, including distinct reflection patterns and reverberation.

If such an RIR is available, it becomes possible to make the room audible without being in it. Any sound such as speech or music can be processed with the RIR and played back over headphones, creating the impression that it occurred inside that room. However, this illusion quickly breaks down if only a single RIR is used. In this case, the same signal is presented to both ears, and the listener has no sense of where the sound is coming from, it is typically perceived *inside the head*. The result resembles the timbre of the room, but it lacks spatial information.

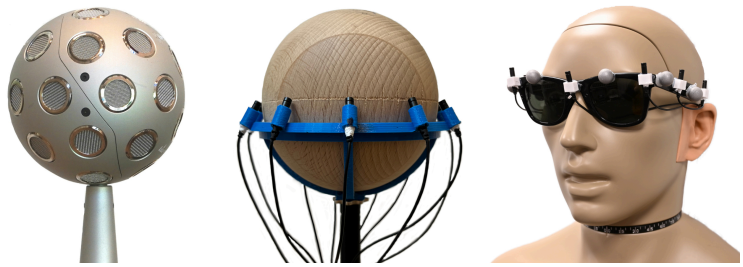


Figure 1.1: Microphone arrays suitable to capture SRIRs. From left to right: a spherical microphone array (Eigenmike EM32), an equatorial microphone array, and a head-worn microphone array on a pair of sunglasses (worn by an artificial head).

To accurately reproduce how a room sounds to a listener, directional information must be captured as well. This leads to the concept of a spatial room impulse response (SRIR). Instead of using a single microphone, an SRIR is measured with multiple microphones arranged in space, allowing directional properties of the arriving sound to be captured. Figure 1.1 shows examples of such microphone arrays, ranging from a spherical microphone array (SMA) to a wearable array on sunglasses. When processed appropriately, an SRIR enables a convincing reproduction of a room's acoustics, including the perception of direction and spatial impression.

The act of making a sound audible in a space where it did not physically occur is known as *auralization*. By analogy to visualization, auralization renders sound rather than images. More formally, it refers to *the process of rendering audible, by physical or mathematical modeling, the sound field of a source in space, in such a way as to simulate the binaural listening experience at a given position in the modeled space* [1]. While auralization focuses on the generation and rendering of signals, the broader concept of creating interactive acoustic environments, in which sources, listeners, and propagation conditions may dynamically change, is also called *virtual acoustics*, a term that is further motivated by recent applications in virtual reality (VR) and augmented reality (AR) and appears in the title of this thesis.

Applications Beyond the conceptual appeal, auralization and virtual acoustics have very practical applications. Using SRIRs, sounds can be placed in real spaces even if they were never physically present there, allowing listeners to experience these environments without being there themselves. A voice, a musical instrument, or any other sound can be made to appear as if it originated in a specific position in a particular room. This capability is useful in acoustic planning, remote communication, perceptual research, and VR/AR.

Throughout this thesis, these ideas are motivated using the AR or telepresence setting in Figure 1.2. A person speaks inside a room, and their voice is captured by a

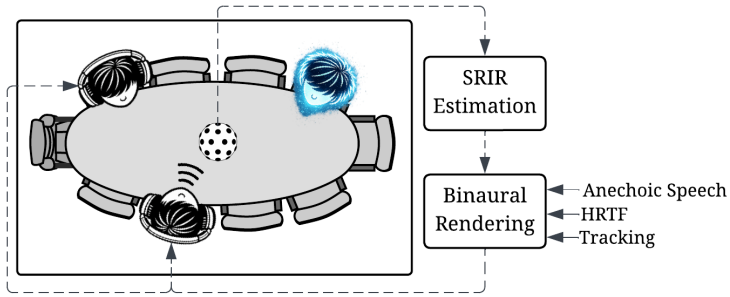


Figure 1.2: An AR or telepresence system is realized by estimating an SRIR from microphone array signals and rendering a virtual sound source (illustrated in blue) whose acoustics match the real environment over headphones.

microphone array. From these recordings, the directional acoustic properties of the room are estimated as an SRIR, applied to an anechoic source signal, and converted to binaural ear signals for headphone playback. The binaural rendering requires a set of head-related transfer functions (HRTFs) to provide localization cues, tracking information to consider movement of the listeners, and an anechoic signal as the virtual source signal. In a telepresence scenario, the source signal originates from a remote participant whose speech has been processed to remove the acoustic influence of their own environment. When played back, the listeners perceive the remote voice as if it were coming from a specific location within the local room. To preserve other sounds in the real environment, such as the speech of the on-site participants or additional sound sources, such systems rely on acoustically transparent (*open*) headphones [2–4].

Structure and Contributions This thesis addresses two tightly coupled challenges in virtual acoustics for practical AR and telepresence applications. The first is the estimation of SRIRs from sensor data in scenarios where dedicated acoustic measurements are unavailable and microphone arrays are typically compact, wearable, and in motion. The second is the perceptually robust binaural rendering with explicit consideration of microphone array characteristics and properties of the SRIR. An overview of the structure of the thesis and the relationship of the included papers is provided in Figure 1.3.

Chapter 2 provides the necessary background, introducing SRIRs, spherical and circular harmonics, and microphone array signal models in a more systematic manner. Building on this foundation, Chapter 3 focuses on the estimation of SRIRs from microphone array signals, particularly introducing an approach based on an estimated reference signal and investigating estimation with moving microphone arrays. Chapter 4 then turns to the rendering stage, discussing contributions to binaural reproduction from array recordings and methods for decomposing SRIRs into direct and diffuse sound

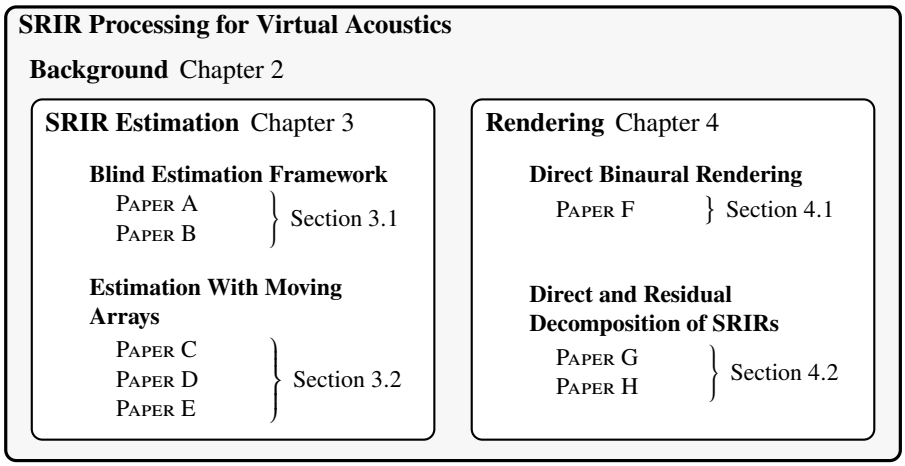


Figure 1.3: Overview of the structure of this thesis and how the papers contribute to the topics.

components. The thesis concludes with Chapter 5, which summarizes the main findings and outlines directions for future research. Finally, Chapter 6 provides an overview of the appended publications and details the specific contributions of each paper.

The contributions of the appended papers can be categorized into SRIR estimation and rendering as shown in Figure 1.3. Paper A to Paper E deal with the estimation of SRIRs. Paper A introduces a modular estimation framework and Paper B provides a comprehensive evaluation of the framework with a head-worn microphone array under stationary conditions. Paper C to Paper E address SRIR estimation from moving arrays, progressing from controlled laboratory configurations to increasingly realistic scenarios. Specifically, Paper C considers rotating arrays, Paper D investigates linearly moving arrays, and Paper E evaluates freely moving head-worn arrays in a more realistic setup utilizing the full estimation framework.

Paper F to Paper H focus on virtual acoustic rendering. Paper F proposes an array-aware extension of the magnitude-least-squares binaural rendering method. Paper G and Paper H describe two different approaches to decompose SRIRs into a direct part and a residual, which is beneficial for parametric rendering approaches.

The contributions of this thesis can be summarized at a conceptual level as follows:

- A general signal model for arbitrary microphone arrays that explicitly accounts for array translation and rotation, and separates the contributions of the acoustic environment and the microphone array.
- A modular framework for SRIR estimation from microphone array recordings,

reformulating blind system identification as an informed estimation problem through the use of a pseudo-reference signal.

- Methods for SRIR estimation with moving microphone arrays, demonstrating that array motion can be compensated for and exploited to increase spatial resolution and extend the effective bandwidth of sound-field representations.
- A unified perspective on direct binaural rendering, clarifying relationships between Ambisonic rendering, beamforming-based binaural reproduction, binaural signal matching, and end-to-end magnitude least squares rendering.
- An array-aware magnitude-least-squares binaural renderer, explicitly accounting for array characteristics to reduce spatio-spectral coloration.
- Two methods for separating direct and residual components of SRIRs, enabling robust extraction and resynthesis of direct sound and early reflections while preserving the directional characteristics of the residual reverberation.

Although not the focus of this thesis, many of the contributions were evaluated perceptually with listening experiments.

Chapter 2

Virtualization of Real-World Acoustics

To auralize a sound source virtually in a real environment via headphone playback, like in the introductory example from Figure 1.2, the rendered signal must convey both head-related cues and room acoustic cues. Head-related cues enable the perception of a sound source in a distinct direction and are essential for localization. Their linear, time-invariant characteristics are captured for an individual listener in the head-related impulse response (HRIR), or their frequency-domain counterpart, the HRTF. Room acoustic cues, on the other hand, ensure that the acoustic environment of the virtual source matches the real-world acoustics so that the source naturally blends in. The linear, time-invariant directional room acoustic cues are captured in the SRIR.

Perceptual Validity of BRIR-Based Virtual Acoustics The combination of HRIR and SRIR is called the binaural room impulse response (BRIR). It captures the linear, time-invariant response to an acoustic impulse from the source to the listener's ear canals. The BRIR can be measured or simulated and is specific to a particular environment, listener, and source-receiver position. The conceptually most straightforward way to obtain a BRIR is through direct measurement using microphones placed in the listener's ear canals. Convolving a source signal with such a BRIR then allows for the binaural reproduction via headphones. Rendering with individually measured BRIRs can be perceptually *authentic*, meaning that it cannot be reliably distinguished from a corresponding real sound source in a direct comparison [5].

In practical scenarios, however, individual BRIR measurements are often infeasible. Instead, measurements are commonly performed using a head and torso simulator (HATS). This approach has been shown to enable *plausible* virtual source rendering,

Figure 2.1: Virtual acoustic rendering can be achieved with a BRIR, for example obtained with a HATS (left). A more flexible approach uses an SRIR captured with a microphone array (center). The renderer then transforms the SRIR into a corresponding BRIR for any listener and arbitrary head rotation, provided that a set of anechoic HRIRs of that listener (top right) and anechoic array transfer functions (ATFs) (bottom right) are available.

defined as *a simulation in agreement with the listener's expectation towards a corresponding real event* [6]. In this context, plausibility refers to the agreement with the listener's internal reference, without a direct comparison to a real sound source. Plausible rendering is therefore a typical target for VR applications.

Yet, recent studies have emphasized the need for a stricter perceptual criterion in applications related to AR, where virtual and real sound sources coexist [7, 8]. In such scenarios, a virtual source is termed *transfer plausible* if it is *believed to be real in the presence of real sound sources* [9]. In contrast to authentic rendering, the virtual source is not compared to a corresponding real source at the same position, but rather to other real sources that differ in signal content and spatial location.

Flexible Capture and Rendering Greater flexibility in capture and rendering is achieved by replacing BRIRs with dynamic rendering approaches that combine HRIRs and SRIRs as illustrated in Figure 2.1. HRIRs can then be individualized and exchanged for different listeners, SRIRs can be manipulated to allow for rotations or translations within the sound field, and SRIR-based representations naturally facilitate the use of microphone arrays integrated into wearable devices. Several such methods have been shown to achieve highly plausible renderings [10–12]. These advantages directly motivate the work presented in this thesis and the corresponding processing pipeline consisting of SRIR estimation and subsequent binaural rendering in Figure 1.2.

While SRIRs could in principle be rendered over a loudspeaker array in an anechoic environment to create virtual acoustic scenes, this thesis focuses on the more practically relevant case of binaural headphone rendering. As outlined in the introductory example, practical AR applications are likely to rely on open headphones that minimally distort real sounds while enabling the presentation of virtual sources. Binaural rendering

then involves combining SRIRs and HRIRs in a dynamic manner, for instance, to account for head rotations based on tracking information. Extensions to six degrees of freedom (6DoF) listener motion have also been proposed [13–16].

Room Acoustic Parameters As an alternative to explicitly estimating SRIRs, some approaches aim to estimate perceptually relevant acoustic parameters and either synthesize SRIRs from them or employ more efficient artificial reverberators such as feedback delay networks (FDNs) [17–19]. Ideally, a set of (directional) room acoustic parameters and corresponding plausibility thresholds could be established through perceptual experiments, such that meeting all thresholds would guarantee a plausible rendering. To date, however, neither a definitive parameter set nor universal thresholds have been identified. The perceptual relevance of individual parameters depends strongly on the experimental setup and the rendering method, and parameters are generally not perceptually independent so that mismatches across several parameters tend to interact and jointly affect perception [20, 21].

Nevertheless, certain parameters have consistently been shown to be important for (transfer) plausible rendering. Most notably, these include the reverberation time (RT) and early-to-late energy metrics such as the direct-to-reverberant energy ratio (DRR) and the clarity. A comprehensive overview of perceptually relevant parameters, experimental findings, and remaining challenges is provided in [20].

Consequently, many practical algorithms focus on estimating RT, DRR, and clarity. A representative overview of classical signal processing approaches was provided by the acoustic characterization of environments (ACE) challenge in 2015 [22], whose methods remain meaningful baselines to this day. More recent work increasingly leverages advances in deep learning, for example [23–28].

Also SRIRs can be used as a basis for deriving room acoustic parameters. Given their perceptual relevance, it is therefore important to assess the accuracy of such parameters when computed from estimated rather than measured SRIRs. This aspect is considered in Paper A, Paper B, and Paper E that focus on the SRIR estimation problem.

The following sections introduce key concepts required for the contributions of this thesis. First, SRIRs are discussed in more detail, followed by an introduction to spherical harmonics (SHs) and circular harmonics (CHs). Finally, a microphone array signal model is presented, which provides a unifying for the methods developed throughout the remainder of the thesis.

2.1 Spatial Room Impulse Responses

In the introduction of this chapter, it was established that SRIRs contain the room acoustic information required for the plausible auralization of virtual sound sources. This section introduces SRIRs more formally, discusses their defining properties and common modeling assumptions, and outlines how they can be obtained either from microphone array measurements or through acoustic simulation.

The term *spatial room impulse response* is not used consistently throughout the literature. In this thesis, an SRIR is defined as *a multichannel room impulse response that captures the directional response of an environment for a single source-receiver position pair*. Since SRIRs describe the acoustic response at a single receiver position, they are typically measured using compact microphone arrays and can be expressed either directly as multichannel array response or using a SH representation. In contrast, distributed microphone arrays can capture directional information over larger spatial extents, but do not directly support directional analysis or rendering for a single, well-defined receiver position.

Some authors restrict the term SRIR to responses expressed in the SH (or Ambisonic) domain and refer to array-domain measurements more generally as multichannel RIRs. Others prefer the term *directional room impulse response* to emphasize the validity of the representation at a single receiver position. However, although strongly limited in practice as will be discussed in Section 3.2, an SRIR may also represent the sound field within a spatial region around the receiver. For this reason, and because it is more widely used, the term SRIR is adopted throughout this thesis.

BRIRs measured with a HATS also fall under this definition of SRIRs. In this case, auralization is directly achieved by convolving a source signal with the measured responses.

Common Modeling Assumptions RIRs, including SRIRs, are typically described as consisting of three components with distinct physical and perceptual characteristics: the direct sound, early reflections, and late reverberation. Figure 2.2 illustrates this for a single-channel RIR.

The direct sound corresponds to the sound traveling along the shortest path from the source to the receiver and, when a direct line of sight exists, appears as the earliest and usually strongest peak in the impulse response. It is commonly modeled as a single delayed and attenuated impulse. Early reflections are produced by a small number of low-order reflections and are typically described explicitly as individual echoes with distinct delays, amplitudes, and directions of arrival [29]. Although often modeled as impulses, in practice, the reflections exhibit frequency-dependent spectral characteristics that lead to temporal spreading due to reflection and absorption at boundaries. Even the direct sound deviates from an ideal impulse as a result of

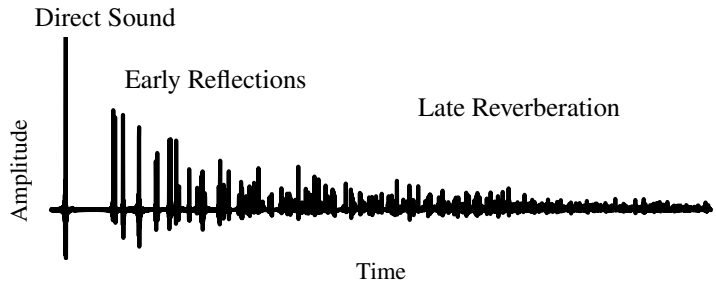


Figure 2.2: Room impulse responses are commonly modeled as consisting of the direct sound, early reflections, and late reverberation.

diffraction and scattering at the microphone baffle.

As time increases, the number of possible reflection paths grows exponentially with reflection order, leading to a rapidly increasing reflection density [30, 31]. The resulting dense superposition of high-order reflections gives rise to the late reverberation, whose fine temporal and spatial structure is difficult to resolve individually and is therefore often approximated as a diffuse random process beyond the mixing time [32]. In practice, this component is often modeled using noise with an exponentially decaying envelope [33, 34].

Although the boundary between early reflections and late reverberation is inherently ambiguous because no strict temporal separation exists, this conceptualization is nevertheless useful for deriving perceptually meaningful room-acoustic parameters and for designing signal processing algorithms that treat the individual components appropriately. Both the SRIR estimation framework in Section 3.1 and the rendering approaches described in Section 4.2 are based on this conceptual model.

Capturing SRIRs Capturing SRIRs in real-world environments requires the use of compact microphone arrays as shown in the introduction in Figure 1.1. Traditionally, SMAs have been employed for this purpose, and a rich body of analytical models exists that facilitates the processing of the recorded signals [35]. In particular, such processing commonly relies on SH representations. However, many practical applications cannot rely on dedicated SMAs and instead must use microphone arrays integrated into existing devices, such as laptops, smartphones, smart speakers, or wearable devices including head-mounted displays, smartglasses, or wristbands. Thus, wearable microphone arrays are explicitly considered in several contributions of this thesis. Their continuous motion, however, introduces additional challenges for SRIR estimation and representation, which are addressed in Section 3.2.

While this thesis focuses on SRIR-based auralization, acoustic simulation is a valid alternative. Accurately simulating SRIRs of a physical environment that are suitable

for plausible auralization remains challenging, as such simulations typically require detailed knowledge of the geometry and frequency-dependent material properties of the space [36]. Nevertheless, recent work has demonstrated that highly plausible renderings, comparable to Ambisonic rendering from measured SRIRs, can be achieved when frequency-dependent reverberation times derived from measured BRIRs are used to calibrate the simulation [12]. Estimated SRIRs, as discussed in Chapter 3, could therefore also serve as input to room acoustic simulations, potentially in combination with geometric information obtained from visual sensing.

For array-independent processing and rendering, SRIRs are often represented using coefficients of SHs or CHs. The following section motivates this representation, introduces SHs and CHs more formally, and discusses their role in the contributions presented in this thesis.

2.2 Spherical and Circular Harmonics

SHs are orthogonal basis functions on the sphere and are of particular importance in audio and acoustics because they provide a compact and physically motivated way to describe how a sound field varies with direction. In spherical coordinates, SHs constitute the angular component of the solution to the acoustic wave equation [37]. Together with appropriate radial basis functions, SH expansions can be used to describe any source-free region of a sound field, a property that will be exploited Section 2.3.

From a signal processing perspective, this makes SHs especially well suited for microphone array processing [35, 38], sound-field estimation and synthesis [39, 40], and spatial audio [41]. They provide a direction-continuous representation that allows sound fields to be rotated, interpolated, and otherwise manipulated in a mathematically well-defined manner. Moreover, the representation is hierarchical: increasing the maximum SH order systematically increases the achievable angular resolution, while truncation directly limits spatial detail.

One of the most prominent applications of SHs in audio is the Ambisonics framework [41–43], which represents sound fields at a single receiver position using an SH decomposition and enables flexible capture, processing, and rendering independent of the microphone array and reproduction format. Within Ambisonics, sound scenes may be encoded either from discrete sound sources or from microphone array recordings, most commonly using SMAs. The resulting SH representation is largely device independent, as array-specific characteristics are compensated for during the encoding stage. This enables flexible downstream processing and rendering, allowing the same Ambisonic signal to be rendered to different loudspeaker layouts or to binaural headphone signals. This decoupling of capture, processing, and rendering also brings practical advantages for consumer applications by facilitating processing routines that support different device generations and user-individualized rendering. Moreover, the

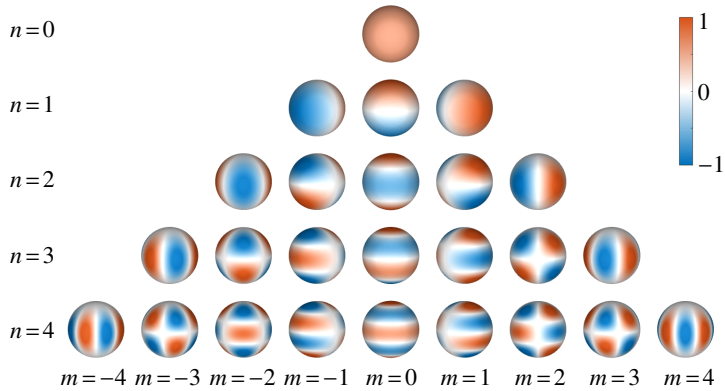


Figure 2.3: Normalized real-valued SHs up to order $N = 4$ in the triangular (n, m) layout.

direction-continuous representation provided by SHs greatly simplifies sound-field rotations, which is essential for efficient dynamic binaural rendering during head movements.

Formally, SHs form a complete and orthogonal basis for square-integrable functions $f(\Omega)$ defined on the surface of a sphere [37]. A commonly used real-valued definition of the SHs is given by

$$Y_n^m(\Omega) = N_n^m \, P_n^{|m|}(\cos(\theta)) \begin{cases} \sqrt{2}\sin(|m|\phi), & m < 0 \\ 1, & m = 0 \\ \sqrt{2}\cos(|m|\phi), & m > 0, \end{cases} \tag{2.1}$$

where $n \in \mathbb{N}_0$ denotes the order, $-n \leq m \leq n$ the degree, $\Omega = \{\phi, \theta\}$ comprises azimuth angle ϕ and zenith (colatitude) angle θ, $P_n^m(\cdot)$ are the associated Legendre polynomials, and

$$N_n^m = (-1)^m \sqrt{\frac{2n+1}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}} \tag{2.2}$$

is a normalization term.

Figure 2.3 illustrates the real-valued SHs up to order $N = 4$. With increasing order n, the basis functions exhibit increasingly fine angular structure, effectively partitioning the sphere into smaller regions.

SH Transform Due to the orthogonality of SHs, any such function $f(\Omega)$ can be expressed as a weighted sum of SHs $Y_n^m(\Omega)$ with corresponding expansion coefficients

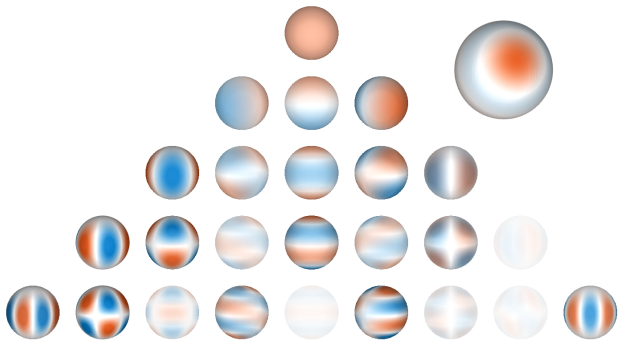


Figure 2.4: The function on the sphere on the top right is expressed via coefficients of SHs. The opacity of the depicted SHs illustrates the magnitude of the corresponding SH coefficient to represent the function.

f_n^m [37],

$$f(\Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} f_n^m Y_n^m(\Omega). \tag{2.3}$$

This equation is also referred to as the *inverse SH transform*. The *SH transform*, allowing to obtain coefficients f_n^m for a given function $f(\Omega)$ is given by

$$f_n^m = \int_{\mathbb{S}^2} f(\Omega)\, Y_n^m(\Omega)\, \mathrm{d}\Omega\,, \tag{2.4}$$

where $\int_{\mathbb{S}^2} \mathrm{d}\Omega = \int_0^{2\pi} \int_0^{\pi} \sin\theta \, d\theta \, d\phi$ denotes the integral over the unit sphere.

Figure 2.4 illustrates how the function in the top right can be represented as a weighted sum of SHs. The weight associated with each SH basis function, its SH coefficient, is visualized by the opacity of the corresponding SH in the figure. Computing these coefficients (the opacity values) from the function in the top right corresponds to applying the SH transform to the function, whereas reconstructing the function from the coefficients requires the inverse transform. Because SHs exhibit progressively finer angular structure with increasing order n, expanding a function using SHs up to a given order yields a directional resolution that increases as higher-order terms are included.

In theory, an infinite number of SHs is required for an exact representation of functions with arbitrarily fine angular structure. In practice, SH expansions are truncated to a finite maximum order N, which directly limits the achievable angular resolution. The inverse discrete SH transform with coefficients $\boldsymbol{f}_N \in \mathbb{C}^{(N+1)^2}$ up to maximum order N is then given by truncating the infinite sum in (2.3),

$$\boldsymbol{f}_Q = \boldsymbol{Y}_Q \boldsymbol{f}_N, \tag{2.5}$$

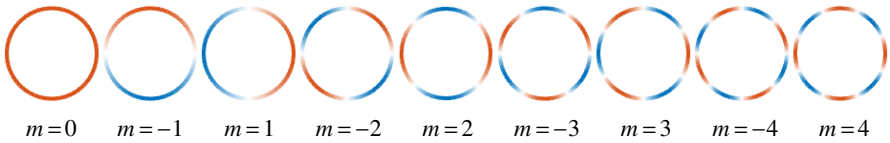


Figure 2.5: Normalized real-valued CHs up to $|m| = 4$. The same colorscale as in Figure 2.3 applies.

where $\boldsymbol{Y}_Q \in \mathbb{R}^{Q\times(N+1)^2}$ contains the SHs evaluated in Q directions. The least squares (LS)-optimal estimate for the discrete SH transform is [35]

$$\boldsymbol{f}_N = (\boldsymbol{Y}_Q^\top \boldsymbol{Y}_Q)^{-1}\boldsymbol{Y}_Q^\top \boldsymbol{f}_Q. \tag{2.6}$$

Circular Harmonics Functions on the circle can be expressed analogously using CHs. In acoustics, they describe two-dimensional sound fields when combined with appropriate radial functions [44]. A real-valued definition of the CHs is given by

$$C_m(\phi) = \begin{cases} \sqrt{2}\sin(|m|\phi), & m < 0 \\ 1, & m = 0 \\ \sqrt{2}\cos(|m|\phi), & m > 0, \end{cases} \tag{2.7}$$

with $m \in \mathbb{Z}$.

CHs up to order $N = 4$ ($|m| \leq 4$) are shown in Figure 2.5. The shown CHs are, up to normalization, equivalent to a horizontal cross-section (at the equator) of the corresponding SHs with the same degree m, located at the edge of the SH triangle in Figure 2.3 (where $n = |m|$).

CHs are of practical relevance because, for a given maximum order N (corresponding to the maximum $|m|$), they provide the same azimuthal resolution as SHs while requiring significantly fewer basis functions, specifically $2N + 1$ instead of $(N + 1)^2$. In microphone array processing, CH processing thus can achieve the same azimuthal accuracy with fewer sensors. Consequently, CHs are particularly beneficial when spatial sampling is restricted to a small number of sensors arranged in a (horizontal) plane, where SH processing becomes ill-conditioned. In this thesis, they are of specific relevance when using wearable microphone arrays of only a small number of microphones arranged approximately on a circle. While many of the following formulations use SHs, they typically apply in the same way to two-dimensional sound fields by replacing the SH terms (and corresponding radial terms) with corresponding CH terms.

Several contributions of this thesis employ SHs or CHs as part of their underlying signal model. Building on two-dimensional sound-field representations, Paper C, Paper D, and Paper E use them to describe signals captured by moving microphone

arrays. Paper F employs an SH-based model for stationary microphone arrays and exploits it within and without the Ambisonics framework. Finally, Paper G uses an SH-based array model to improve the extraction of reflections from SRIRs by more accurately accounting for the array influence. The following section introduces a general microphone array signal model that serves as the foundation for these contributions.

2.3 Microphone Array Signal Models

Several contributions in this thesis rely on signal models to derive optimal processing routines. This section introduces a general microphone array signal model that makes the separation between room-dependent and array-dependent effects explicit and provides the foundation for several of the proposed SRIR estimation and rendering methods.

Consider a single sound source in a room observed by a microphone array of M microphones. In the frequency domain, a commonly used narrowband signal model for the array signals $\boldsymbol{x} \in \mathbb{C}^M$ at frequency f is

$$\boldsymbol{x}(f) = \boldsymbol{v}(f)\, s(f) + \boldsymbol{n}(f), \tag{2.8}$$

where $\boldsymbol{v} \in \mathbb{C}^M$ contains the acoustic transfer functions between the source and the individual microphones, s is the source signal, and $\boldsymbol{n} \in \mathbb{C}^M$ is additive noise. Each entry of $\boldsymbol{v}$ implicitly combines the spatial room transfer function representing the effects of sound propagation in the room and the spatial transfer function of the microphone array, depending on array geometry, surface properties, and microphone directivities.

A partial separation between room- and array-dependent contributions can be achieved by selecting a reference microphone with corresponding transfer function v_1 and introducing relative transfer functions (RTFs) [45]

$$\tilde{\boldsymbol{v}}(f) := \left[1, \quad \frac{v_2(f)}{v_1(f)}, \quad \ldots, \quad \frac{v_M(f)}{v_1(f)}\right]^\top, \tag{2.9}$$

such that $\boldsymbol{v} = v_1\, \tilde{\boldsymbol{v}}$. Substituting this factorization into (2.8) yields

$$\boldsymbol{x}(f) = v_1(f)\, \tilde{\boldsymbol{v}}(f)\, s(f) + \boldsymbol{n}(f). \tag{2.10}$$

With this factorization, $\tilde{\boldsymbol{v}}$ captures only the relative spatial structure of the sound field across the array, while the transfer function from the source to the reference microphone v_1 absorbs all components that are common across microphones. It is typically assumed that the dominant room-dependent contributions are contained in v_1. However, any spatial characteristics of the room remain embedded in $\tilde{\boldsymbol{v}}$, together with the spatial response of the array.

Although this formulation is useful for many array processing tasks [46–50] and provides partial robustness to changes in source and room conditions, it does not provide a strict separation between spatial room and array contributions. This limits its usefulness for SRIR estimation and rendering, where the goal is to obtain a directional representation of the room acoustics that is independent of the specific microphone array used for recording.

For notational clarity, the noise term is omitted in the following. The resulting algorithms remain optimal under the common assumption of spatially white Gaussian noise but, in practice, regularization is necessary to obtain stable and robust solutions in the presence of finite data, noise, and model imperfections [51].

Separation of Room and Array Contributions To achieve such a separation, we reformulate the problem from a sound-field perspective and assume that the microphone array is contained in a source-free region of the sound field. In this region, the sound field can be represented as a continuous superposition of plane waves. Temporarily neglecting the influence of the microphone array, the sound pressure at a position $\boldsymbol{r}$ is given by [35]

$$p(f, \boldsymbol{r}) = s(f) \int_{\mathbb{S}^2} a(f, \Omega)\, \mathrm{e}^{\mathrm{i}k\boldsymbol{r}^\top \boldsymbol{u}(\Omega)} \mathrm{d}\Omega, \tag{2.11}$$

where k denotes the wavenumber and $\boldsymbol{u}(\Omega)$ is the unit vector pointing in the plane-wave arrival direction Ω. The function $a(f, \Omega)$ is the plane wave density associated with the room transfer function and fully characterizes how the room maps the source signal to a directional sound field.

Assuming the microphone array is centered in the coordinate origin $\boldsymbol{r} = \boldsymbol{0}$, the sound pressure in direction Ω is given by

$$p(f, \Omega) = a(f, \Omega)\, s(f). \tag{2.12}$$

Now considering the array's influence, the signal at microphone l can be expressed as the directional superposition of the incident sound field weighted by the lth microphone's directional response d_l,

$$x_l(f) = \int_{\mathbb{S}^2} d_l(f, \Omega)\, p(f, \Omega)\, \mathrm{d}\Omega. \tag{2.13}$$

When considering wearable microphone arrays, the directional array response d_l is also influenced by the interaction of sound waves with the wearer's body. In this context, d_l has also been termed wearable-device related transfer function (WDRTF) [52].

Stacking all directional array transfer functions into a vector $\boldsymbol{d} = [d_1, \ldots, d_M]^\top$ yields the array signals

$$\boldsymbol{x}(f) = \int_{\mathbb{S}^2} \boldsymbol{d}(f, \Omega)\, p(f, \Omega)\, \mathrm{d}\Omega. \tag{2.14}$$

To obtain a discrete coefficient-based model, the plane wave density is expanded into SHs via (2.3),

$$a(f,\Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} a_n^m(f)\, Y_n^m(\Omega). \tag{2.15}$$

Substituting this expansion into the array measurement equation (2.14) and exchanging summation and integration yields

$$\boldsymbol{x}(f) = s(f) \sum_{n=0}^{\infty} \sum_{m=-n}^{n} a_n^m(f) \int_{\mathbb{S}^2} \boldsymbol{d}(f,\Omega)\, Y_n^m(\Omega)\, \mathrm{d}\Omega. \tag{2.16}$$

By defining the columns of the spatial array response matrix $\boldsymbol{D} \in \mathbb{C}^{M\times(N+1)^2}$ as

$$\boldsymbol{D}_{(:,n,m)}(f) := \int_{\mathbb{S}^2} \boldsymbol{d}(f,\Omega)\, Y_n^m(\Omega)\, \mathrm{d}\Omega, \tag{2.17}$$

limiting the SH coefficients a_n^m to a maximum order $n \leq N$ and stacking them into a vector $\boldsymbol{a} \in \mathbb{C}^{(N+1)^2}$, the signal model can be written compactly as

$$\boxed{\boldsymbol{x}(f) = \boldsymbol{D}(f)\, \boldsymbol{a}(f)\, s(f).} \tag{2.18}$$

In this formulation, $\boldsymbol{a}$ represents the spatial room transfer function in the form of SH coefficients of the plane wave density. The spatial array response matrix $\boldsymbol{D}$ contains the SH coefficients of the directional array response and, by summing over its columns, transforms from the SH domain to the microphone signals. This factorization makes the separation between room acoustics and array influence explicit and provides a basis for array-independent signal representations and algorithms. Following the SRIR definition from Section 2.1, both the SH-domain representation $\boldsymbol{a}$ and the microphone-domain representation $\boldsymbol{Da}$ are considered spatial room transfer functions in this thesis, although only the former is array independent.

Spatial Array Response Matrix For SMAs, analytical expressions for the array response matrix $\boldsymbol{D}$ exist and can be obtained by expressing the sound-field equation (2.11) in spherical coordinates,

$$p(f,r,\Omega) = s(f) \sum_{n=0}^{\infty} \sum_{m=-n}^{n} a_n^m(f) b_n(kr) Y_n^m(\Omega), \tag{2.19}$$

where $b_n = 4\pi \mathrm{i}^n j_n$ are called radial terms and j_n are the spherical Bessel functions [35].

Comparing (2.16) and (2.19) reveals that, for open SMAs (without baffle), the array response matrix is analytically defined by

$$\boldsymbol{D}_{\mathrm{SMA}}(f) = \boldsymbol{Y}_{\mathrm{M}}\, \mathrm{diag}\{\boldsymbol{b}(f)\}\,, \tag{2.20}$$

where $\boldsymbol{Y}_\mathrm{M} \in \mathbb{R}^{M\times(N+1)^2}$ contains the SHs evaluated at the microphone positions and $\boldsymbol{b} \in \mathbb{C}^{(N+1)^2}$ the radial terms evaluated at the array radius. Similar solutions for other spherical microphone configurations such as rigid arrays only differ by the definition of their radial terms and are derived by considering appropriate boundary conditions imposed by the array [37, 53, 54].

For arbitrary microphone arrays, such analytical expressions are generally not available. The array response matrix can, however, be estimated from a set of anechoic ATFs $\boldsymbol{P}_\mathrm{A} \in \mathbb{C}^{M\times Q}$ in a discrete grid of Q directions. These ATFs can be analytically described as microphone signals excited by unit-amplitude plane waves from the corresponding directions [35],

$$\boldsymbol{P}_\mathrm{A}(f) = \boldsymbol{D}(f)\boldsymbol{Y}_Q^\top. \tag{2.21}$$

The spatial array response matrix is then estimated as an optimal weighted-least-squares fit between the model and the measured (or simulated) ATFs [55]

$$\begin{aligned}\hat{\boldsymbol{D}}(f) &= \arg\min_{\boldsymbol{D}} \left\| \left(\boldsymbol{D}\boldsymbol{Y}_Q^\top - \boldsymbol{P}_\mathrm{A}(f) \right) \boldsymbol{W}^{\frac{1}{2}} \right\|_\mathrm{F}^2 \\ &= \boldsymbol{P}_\mathrm{A}(f)\, \boldsymbol{W}\boldsymbol{Y}_Q \left(\boldsymbol{Y}_Q^\top \boldsymbol{W}\boldsymbol{Y}_Q \right)^{-1},\end{aligned} \tag{2.22}$$

where $\boldsymbol{Y}_Q \in \mathbb{R}^{Q\times(N+1)^2}$ contains the SH basis evaluated in the ATF grid directions, and $\boldsymbol{W} \in \mathbb{R}^{Q\times Q}$ is a diagonal matrix of quadrature weights that account for the discrete and potentially non-uniform sampling of the Q directions.

Ambisonic Encoding From Microphone Arrays Based on the signal model in (2.18), Ambisonic encoding can be interpreted as estimating a direction-continuous description $\boldsymbol{a}s$ of the sound field that is independent of the specific microphone array used for capture. Considering a more general multi-source formulation, the vector $\sum_i \boldsymbol{a}_i s_i$ represents the time-varying SH coefficients of the sound scene, while the spatial array response matrix $\boldsymbol{D}$ describes how this SH representation is observed by the microphone signals. Ambisonic encoding therefore aims to invert the array response matrix in order to recover the SH coefficients from the array signals.

For SMAs (2.20), this requires applying the SH transform (2.6) and inverting the radial terms to obtain an array-independent SH representation in the center of the array at $r = 0$ [41],

$$\boldsymbol{E}_\mathrm{SMA}(f) = \mathrm{diag}\{\boldsymbol{b}(f)\}^{-1} \left(\boldsymbol{Y}_\mathrm{M}^\top \boldsymbol{Y}_\mathrm{M}\right)^{-1} \boldsymbol{Y}_\mathrm{M}^\top. \tag{2.23}$$

An explicit encoder is therefore also required for arbitrary arrays to compensate for array effects and recover SH coefficients that represent the sound field at the array center.

Since plane waves form a complete basis for source-free sound fields, an encoder that correctly maps plane waves to their SH representations, by linearity, also provides a valid SH encoding for arbitrary superpositions of plane waves, and thus for general sound scenes within the modeled spatial bandwidth. This leads to a general formulation of an Ambisonic array encoder that, in an LS-optimal way, aims to map the array response to incoming plane waves onto the corresponding ideal SH coefficients [55, 56],

$$\begin{aligned} \boldsymbol{E}_{\mathrm{LS}}(f) &= \arg\min_{\boldsymbol{E}} \left\| \left(\boldsymbol{E}\boldsymbol{P}_{\mathrm{A}}(f) - \boldsymbol{Y}_Q^{\top} \right) \boldsymbol{W}^{\frac{1}{2}} \right\|_{\mathrm{F}}^2 \\ &= \boldsymbol{Y}_Q^{\top} \boldsymbol{W} \boldsymbol{P}_{\mathrm{A}}^{\mathrm{H}}(f) \left(\boldsymbol{P}_{\mathrm{A}}(f) \boldsymbol{W} \boldsymbol{P}_{\mathrm{A}}^{\mathrm{H}}(f) \right)^{-1} . \end{aligned} \tag{2.24}$$

In practice, the inverse in (2.24) requires regularization, and the encoder is typically diffuse-field equalized above the spatial aliasing limit [57–59].

Variants of the signal model from (2.18) form the basis for several estimation and processing methods in this thesis. By replacing the SH basis with CHs, the same signal model can be applied to describe two-dimensional sound-field relations, for example for approximately circular arrays in the horizontal plane. This is done in Paper C, Paper D, and Paper E, where equatorial and head-worn microphone arrays are employed.

Chapter 3

Spatial Room Impulse Response Estimation

As discussed in Section 2.1, SRIRs are a key ingredient for the plausible auralization of virtual sound sources in real-world environments. In practice, however, measured SRIRs are rarely available, as dedicated acoustic measurements are typically infeasible outside laboratory conditions. In the introductory example, a user aims to experience an acoustically matched virtual sound source without providing explicit acoustic parameters or requiring an expert to measure RIRs. Consequently, SRIRs must be estimated from sensor data that are readily available in practical scenarios, most notably microphone (array) recordings of naturally occurring sounds such as speech.

The inverse problem of recovering an RIR from a reverberant signal, modeled as the convolution of an unknown source signal with an unknown RIR, is commonly referred to as *blind* system identification [60]. If the source signal, often termed a *reference*, is known or can be estimated, the problem instead becomes an *informed* or *semi-blind* identification task.

Traditional methods for blind multichannel RIR estimation are largely based on channel cross-relations [61–66]. In a nutshell, these algorithms exploit the observation that the convolution of the signal observed at microphone i with the RIR of microphone j equals the convolution of the signal observed at microphone j with the RIR of microphone i, i.e.,

$$x_i * h_j = x_j * h_i. \tag{3.1}$$

By formulating this relation for all microphone pairs, a system of equations is obtained that can be solved in a LS or least-mean-squares sense.

Historically, such approaches have been developed primarily for communication channel identification, where impulse responses typically consist of only a few taps.

While some works have extended these techniques to acoustic scenarios with several hundred or even up to 1024 taps [66], none have demonstrated robust estimation of acoustic RIRs of realistic duration at typical audio sampling rates, which would require tens of thousands of taps. We investigated this limitation in PAPER A, where we showed that cross-relation-based algorithms fail to converge for such long impulse responses.

In the same work, however, a different strategy was shown to be effective: instead of directly estimating the RIRs, a reference signal termed a *pseudo-reference* is first estimated from the microphone signals. This pseudo-reference can then be used in conjunction with informed system identification techniques to recover the RIRs. The next section reviews this approach in more detail and establishes it as a central building block of the SRIR estimation framework developed in this thesis.

Neural Network Approaches Beyond model-based signal processing methods, neural network approaches have recently gained increasing attention for blind RIR estimation. Many existing works focus on single-channel setups and are often limited in frequency range, yet their results are promising and open up new interesting ways of addressing the problem. Typically, the neural network architecture consists of an encoder that extracts room acoustic features from an audio signal into a compressed latent representation, and a decoder that reconstructs an RIR from this latent space [67, 68]. Recent studies have extended this paradigm in various directions. Some approaches jointly optimize RIR estimation and acoustic matching [69], others focus on generating perceptually valid RIRs [70] or employ physically motivated parametric models [71]. Extensions to multi-source scenarios [72], estimation of RIRs at arbitrary receiver positions [73], multimodal audio-visual learning [74], and joint RIR estimation and dereverberation [75, 76] have also been proposed.

Recent work has also explored alternatives to explicit RIR estimation. One such direction is *acoustic matching*, where the acoustics of one recording are transferred to another by exploiting latent representations learned from data [27, 77, 78]. These approaches bypass explicit RIR estimation altogether and instead aim to directly achieve similarity between recordings. Another complementary line of research seeks to infer room acoustic properties from visual information alone [79–84]. While visual methods can provide valuable prior information about an environment, they typically do not yield a full spatially resolved impulse response suitable for rendering.

When large amounts of realistic training data and computational resources are available, learning-based methods may achieve excellent performance and surpass classical signal processing approaches. However, practical systems for AR or telepresence applications often operate under strict constraints in terms of data availability, computational complexity, and generalization to unseen environments. In such settings, purely data-driven approaches may struggle to provide robust solutions, and a modular estimation framework, such as the one introduced below, may provide an opportunity

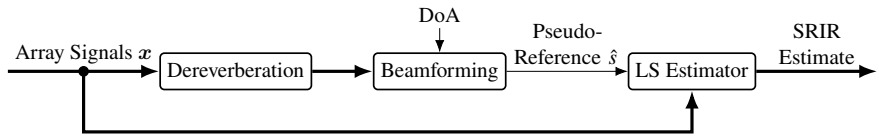


Figure 3.1: The SRIR is estimated from the array signals $\boldsymbol{x}$ and the pseudo-reference $\hat{s}$, which is obtained through dereverberation and beamforming. Bold lines indicate multichannel signals. Figure adapted from PAPER A.

to combine model-driven digital signal processing and deep learning methods in an efficient and high-performing manner in the future.

3.1 Estimation Using a Pseudo-Reference Signal

With traditional blind multichannel system identification methods failing to converge for realistic acoustic RIRs, an alternative strategy is required. In PAPER A, we therefore proposed a modular estimation framework that reformulates the blind problem as an informed one by introducing a pseudo-reference signal. While the general idea of reference estimation has previously been explored in the context of reverberation time estimation [85] and parametrization of SH-domain SRIRs [86], PAPER A was the first to demonstrate the applicability of the full framework to array SRIR estimation.

The framework is motivated by the common conceptual decomposition of RIRs into direct sound, early reflections, and late reverberation (see Figure 2.2). An overview of the signal flow is shown in Figure 3.1. The single-channel pseudo-reference signal $\hat{s}$ is an estimate of the anechoic source signal and is derived from the multichannel array signals. It is obtained by suppressing late reverberation through dereverberation, and extracting the direct sound with minimal distortion while attenuating early reflections via beamforming.

Here, dereverberation is applied prior to beamforming because the employed dereverberation method benefits from multichannel information. However, the order of dereverberation and beamforming can be exchanged if a high-performing single-channel dereverberation method is available. Once the pseudo-reference signal is available, standard estimators like a multichannel Wiener or recursive least squares (RLS) filter can be employed to obtain the SRIR [87]. While the framework is modular and allows for different algorithmic choices within each processing block, the blocks must satisfy certain properties to ensure the validity of the pseudo-reference.

The framework can be applied either directly to array signals or after encoding to Ambisonics. While only the latter yields an array-independent SRIR representation, the former can be combined with an array-aware binaural renderer (Section 4.1) so that the binaural output is again compensated for the array influence.

Dereverberation When dereverberation is applied to the multichannel array signals, it is crucial that inter-channel relationships exploited by the subsequent beamformer are preserved. In particular, the algorithm must not distort the direct sound or alter the relative time differences of arrival between microphones. The generalized weighted prediction error (GWPE) algorithm satisfies these requirements [88]. For each short-time Fourier transform (STFT) subband b and time index τ, GWPE computes a linear prediction matrix $\boldsymbol{G}_b$ that predicts the current multichannel signal vector $\boldsymbol{x}_b(\tau)$ from delayed versions $\boldsymbol{x}_b(\tau - \delta)$ and suppresses late reverberation by subtracting the prediction from the array signals,

$$\hat{\boldsymbol{x}}_b(\tau) = \boldsymbol{x}_b(\tau) - \sum_{\delta=\Delta}^{\Delta+K_b-1} \boldsymbol{G}_b^{\mathrm{H}}(\delta)\, \boldsymbol{x}_b(\tau - \delta)\,. \tag{3.2}$$

A key assumption underlying GWPE is that the source signal exhibits non-negligible temporal correlation only within a short interval of Δ samples, referred to as the prediction delay. Components arriving later than this delay are treated as reverberation and suppressed. For speech signals, Δ is typically chosen between 10 to 40 ms. While GWPE can in principle be applied to other signal types, only a single prediction delay can be specified. For mixtures of signals with different effective correlation lengths, for instance speech and a sustained piano tone, dereverberation will therefore either affect only components arriving after the longest correlation interval, corresponding to the sustained tone and leaving much of the reverberation excited by the speech intact, or suppress signal components with long temporal autocorrelation, effectively cutting off the sustained tone. In such cases, alternative dereverberation approaches, including recent deep learning-based methods trained on signal mixtures, may be more suitable [75, 89].

Beamforming Following dereverberation, beamforming is employed to directionally extract the direct sound component while maximally suppressing signal components arriving from other directions. These undesired components typically correspond to early reflections in single-source scenarios and additionally to interfering sources in multi-source scenarios. A distortionless beamformer can be defined as

$$\boldsymbol{w}_{\mathrm{BF}}(f) = \frac{\boldsymbol{R}_{\mathrm{c}}^{-1}(f)\, \boldsymbol{p}_{\mathrm{A}}(f)}{\boldsymbol{p}_{\mathrm{A}}^{\mathrm{H}}(f) \boldsymbol{R}_{\mathrm{c}}^{-1}(f) \boldsymbol{p}_{\mathrm{A}}(f)}\,, \tag{3.3}$$

where $\boldsymbol{p}_{\mathrm{A}}$ denotes the anechoic ATF corresponding to the desired source direction and $\boldsymbol{R}_{\mathrm{c}}$ is an appropriate spatial covariance matrix. Depending on the choice of $\boldsymbol{R}_{\mathrm{c}}$, this formulation yields a matched-filter beamformer (identity matrix), a minimum-power distortionless response (MPDR) beamformer (signal covariance matrix), or a minimum-variance distortionless response (MVDR) beamformer (noise covariance matrix) [51]. The beamformer requires an estimate of the source direction of arrival (DoA) to select the appropriate ATF. In Paper A and Paper E, we estimated the DoA using the multiple signal classification (MUSIC) algorithm [90].

SRIR Estimation Once the pseudo-reference signal $\hat{s}$ has been obtained, SRIR estimation reduces to an informed system identification problem in which the room transfer function is the only unknown. The spatial room transfer function (the frequency-domain counterpart of the SRIR) is then estimated by minimizing the expected squared error between the filtered pseudo-reference and the microphone array signals. A microphone-domain spatial room transfer function estimate is obtained by

$$\hat{\boldsymbol{w}}(f) = \arg\min_{\boldsymbol{w}} \mathbb{E}\left\{\|\boldsymbol{w}\hat{s}(f) - \boldsymbol{x}(f)\|_2^2\right\}, \tag{3.4}$$

and an array-independent SH- or CH-domain estimate by

$$\hat{\boldsymbol{w}}(f) = \arg\min_{\boldsymbol{w}} \mathbb{E}\left\{\|\boldsymbol{D}(f)\boldsymbol{w}\hat{s}(f) - \boldsymbol{x}(f)\|_2^2\right\}. \tag{3.5}$$

Paper A and Paper B assumed time-invariant conditions and employed a multichannel Wiener filter (MWF), whereas Paper C, Paper D and Paper E addressed scenarios with moving arrays and adopted an RLS formulation. In practical implementations, regularization is necessary to mitigate noise amplification in frequency regions where the pseudo-reference has low energy.

Discussion The studies in Paper A and Paper B demonstrate that the proposed framework enables accurate SRIR estimation in static scenarios across different array geometries, including circular, equatorial, spherical, and head-worn configurations. A central outcome is that dereverberation substantially improves SRIR estimation compared to processing without dereverberation. By attenuating late reverberation during pseudo-reference estimation, dereverberation reduces RT errors and yields DRR estimates that, in the vast majority of cases, fall below the just-noticeable difference (JND).

The applicability of the framework to head-worn microphone arrays integrated into wearable devices further highlights its practical relevance. In such configurations, the estimated SRIRs capture RT and DRR with an accuracy that is superior to dedicated parameter estimators developed in the context of the ACE challenge [22]. Moreover, the parameter estimation remains robust in the presence of noise, interfering speech, and DoA errors. The estimated SRIRs exhibit close agreement with reference responses in both early reflection structure and overall energy decay, as illustrated in Figure 3.2.

Although both time-domain and spectral analysis, as well as parameter estimation performance suggest high accuracy in such static scenarios, the estimated pseudo-reference is inevitably imperfect. DoA errors, limited beamformer directivity, incomplete suppression of late reverberation, noise, and interfering sources introduce imperfections that may manifest in the SRIR estimates as narrowband ringing artifacts. To address this issue, the late reverberation tail can be resynthesized using spectrally shaped, exponentially decaying noise based on octave-band RT, DRR, spatial covariance, and energy decay of the estimated SRIR. Owing to the narrowband nature of the

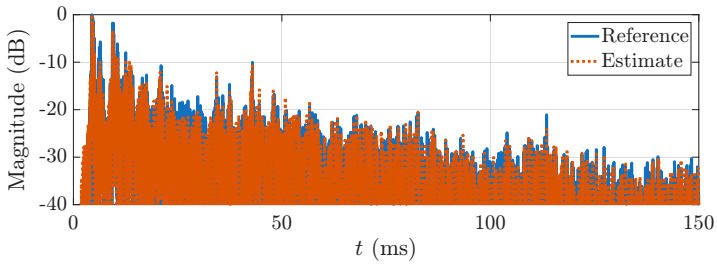


Figure 3.2: One channel of a reference SRIR and an SRIR estimate obtained from speech captured with a microphone array in smartglasses.

artifacts, this simple resynthesis strategy effectively removes the ringing artifacts while preserving perceptually relevant room characteristics. Binaural renderings based on the resynthesized SRIRs then, in some cases, reproduce the perceived room acoustics as faithfully as a measurement, and, in most cases, more faithfully than a measured SRIR of a different but similarly sized room.

Instead of relying on far-field speech, SRIR estimation using the voice of the person wearing the smartglasses may be particularly relevant in practical VR or AR applications using head-worn devices. To support the creation of virtual far-field sources in such scenarios, the corresponding parameter estimates need to be compared to far-field references. In scenarios with a high signal-to-noise ratio (SNR), the blind estimation framework then yields RT estimates that are comparable in accuracy to baseline estimators operating on far-field speech. Such high-SNR conditions are particularly realistic due to the close proximity of the wearer's mouth to the microphone array, and investigations based on anechoic measurements showed that SNRs in scenarios using own speech are typically about 20 dB higher than in far-field speech scenarios.

Naive estimation of DRRs from the wearer's own speech naturally results in significantly higher DRRs than far-field reference values due to the increased direct sound level. However, follow-up work showed that extrapolating these own-speech DRR estimates to far-field DRRs is feasible above the Schröder frequency in sufficiently reverberant rooms [91].

From an application perspective, wearable microphone arrays are of particular interest. However, they introduce an additional challenge in that the array moves continuously with the user. The next section discusses how the proposed SRIR estimation framework can be extended to such dynamic settings.

3.2 Estimation From Moving Arrays

Wearable microphone arrays are particularly valuable for RIR estimation in VR and AR systems, since these systems often already require head-worn devices, such as smartglasses or headsets, for visual tracking and rendering. In the introductory AR example, either the user or another person in the room may be equipped with such an array to capture the speech. Wearable devices, however, introduce a unique challenge: the microphones move continuously during recording. Two key questions then arise: can the array motion be compensated for to obtain accurate SRIR estimates, and, can it be exploited to capture richer directional or spatial information?

While no previous work has specifically addressed SRIR estimation with moving arrays, related research exists in the context of RIR measurement and sound-field estimation [92–98]. The measurement methods typically assume carefully designed measurement signals and the sound-field methods aim to reconstruct the sound field over a large spatial domain. In contrast, the methods introduced in Paper C, Paper D, and Paper E aim to reconstruct the directional sound field at a single receiver location for the purpose of auralization. Although the underlying signal models are related, the methodologies and evaluation criteria differ.

Methodology Paper C, Paper D, and Paper E extend the stationary signal model from (2.18) by explicitly incorporating array translation and rotation and estimating the SRIR as CH-domain sound-field coefficients that are valid in a region around the array's initial position. The array signals at the rotated, translated position are then obtained by applying linear rotation and translation operators to the CH-domain representation,

$$\boldsymbol{x}(f, \alpha, \phi_\mathrm{t}, r_\mathrm{t}) = \boldsymbol{D}(f)\boldsymbol{R}(\alpha)\boldsymbol{T}(f, \phi_\mathrm{t}, r_\mathrm{t})\boldsymbol{a}(f)s(f). \tag{3.6}$$

Here, $\boldsymbol{R}$ represents an azimuthal rotation of the array by angle α, and $\boldsymbol{T}$ models a translation defined by angle ϕ_t and distance r_t. Assuming an estimate of $\boldsymbol{D}$ is available from anechoic ATFs and (2.22), a position tracking system provides α, ϕ_t, and r_t, and a pseudo-reference approximates s, (3.6) can be solved for $\boldsymbol{a}$ with an RLS filter.

The same formulation can be applied to arbitrary 6DoF movements using SHs and their corresponding rotation and translation theorems [99]. However, the number of coefficients required to represent a three-dimensional sound field, $(N + 1)^2$ instead of $2N + 1$ for CHs, and the typical arrangement of microphones mainly around the front and sides of the user's head make this impractical for current head-worn arrays with a limited number of microphones.

Although this concept has not previously been applied to moving microphone arrays, SH and CH translation operations have been used similarly to combine signals from multiple microphone arrays [100–102] and to enable translational head movements during binaural rendering [103, 104].

Discussion While the frequency-independent rotation operator $\boldsymbol{R}$ accurately represents azimuthal rotation for any given maximum order N of the CH coefficients, the accuracy of the translation operation $\boldsymbol{T}$ for a translation distance r is strictly limited by frequency and CH order. As discussed in [104], the translation is only accurately modeled up to $kr < N$ for the zeroth-order CH coefficients, and up to $kr < N - m$ for higher orders m.

The results presented in Paper C and Paper D show that controlled microphone array motion can not only be compensated for but also actively exploited in SRIR estimation. In particular, array rotation provides additional spatial diversity that can be used to improve directional resolution. By combining observations during continuous rotation of a microphone array supporting a maximum CH order N_a, it becomes possible to estimate SRIR coefficients of higher order $N > N_\mathrm{a}$. This allows SRIR estimates obtained from compact rotating arrays to achieve an accuracy comparable to measurements from static arrays with a significantly larger number of microphones.

SRIR estimation during translational array motion presents a significantly greater challenge than rotation alone, since, from the array's perspective, translational movement causes the direct sound and early reflections to change not only in direction but also in time. As a result, the achievable accuracy of SRIR estimates is generally lower compared to SRIRs measured with static arrays. Nevertheless, accurate estimates can still be obtained at low and mid frequencies, and combining multiple translated observations in an optimal LS sense allows the effective CH order N of the SRIR estimate to exceed the array's native limit N_a.

This is illustrated for a single point source in Figure 3.3, where sound-field reconstruction accuracy is used as a proxy for CH coefficient accuracy. This concept follows directly from (2.19), which states that the sound pressure at any point within a source-free region can be reconstructed from a complete set of sound-field coefficients. When only a finite number of coefficients is available, accurate reconstruction is confined to the region $kr \leq N$ [105]. Conversely, if the reconstructed field remains accurate up to a given value of kr, this indicates that the corresponding CH coefficients are accurate up to order N. Figure 3.3 shows that, by combining multiple observations of order N_a and estimating CH coefficients of order $N > N_\mathrm{a}$, the resulting SRIR coefficients remain accurate up to $kr = N$, extending the valid reconstruction region beyond the native limit $kr = N_\mathrm{a}$.

The full blind estimation framework from Figure 3.1, including adaptive dereverberation and beamforming, was applied to a head-worn microphone array embedded in sunglasses (see Figure 1.1) and worn by a human participant performing natural head movements under more realistic conditions in Paper E. In informed settings, where access to a reference signal is available, acoustic parameter estimates then remain comparable to those obtained in stationary scenarios. However, when based on the estimated pseudo-reference, performance degrades, and explicitly modeling the array motion does not provide a benefit anymore over estimating an omnidirectional RIR

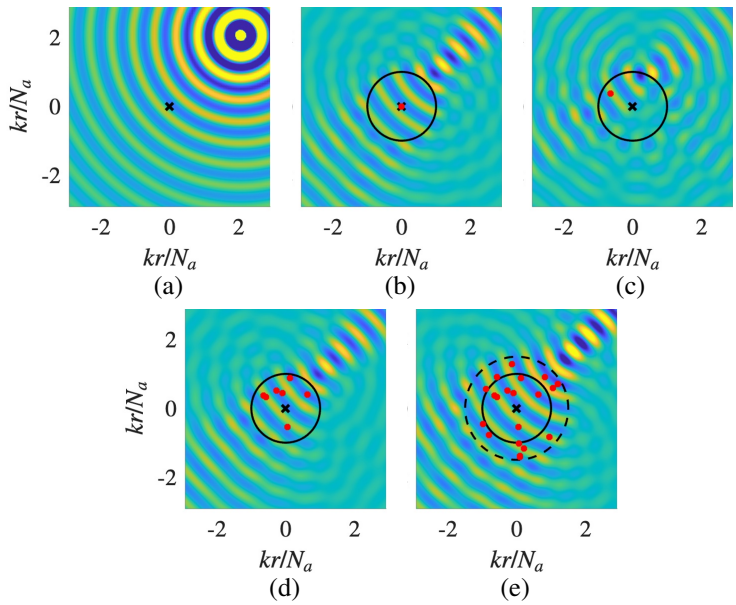


Figure 3.3: The real part of the sound pressure of a point source in (a) is accurately reproduced within $kr = N_\mathrm{a}$ (black circle) in (b), using CH coefficients of order N_a obtained from a microphone array (red dot) at the origin (black cross). In (c), CH coefficients measured by an off-origin array are translated to the origin, but the resulting sound field is inaccurate for $kr < N_\mathrm{a}$. In (d), combining multiple off-origin observations of order N_a yields accurate origin coefficients for $kr < N_\mathrm{a}$. Finally, in (e), with a sufficient number of N_ath-order observations, the coefficient order can be increased to $N > N_\mathrm{a}$, enabling accurate field reproduction within $kr < N$ (dashed circle). Figure adapted from Paper D.

using the zeroth-order CH coefficient only.

While room acoustic parameters such as RT and DRR can still be estimated with promising accuracy under these conditions, limitations remain in the estimation of the direct sound magnitude and the directional energy distribution. As a consequence, the estimated SRIRs are not directly suitable for perceptual evaluation and must be resynthesized using an idealized direct sound impulse combined with isotropic diffuse reverberation shaped to match the estimated parameters. Listening experiments based on these responses indicate that motion-aware SRIR estimation does not consistently provide perceptual advantages over the simpler omnidirectional model. In particular, modeling natural head-worn motion using low-order CH representations obtained from a practical microphone array appears to introduce perceptually relevant errors. Nevertheless, omnidirectional estimation using a zeroth-order CH coefficient remains practically useful, as renderings based on such an RIR are perceived as closer to a

measurement-based reference than renderings from measurements conducted in other rooms.

Chapter 4

Binaural Spatial Room Impulse Response Rendering

Binaural rendering of a virtual sound source generally consists of applying the desired acoustic characteristics to a dry source signal and converting the result into binaural ear signals. In the conceptually straightforward case, where a BRIR is available that contains both head-related and room-acoustic cues, the rendering is simply achieved by convolution of BRIR and signal [106]. In more general settings, such as when using SRIRs, these two aspects must be handled separately, such that the source signal is reverberated and rendered binaurally in two separate steps.

Methods for binaural rendering of virtual acoustics can broadly be classified into direct and parametric approaches [18]. Direct methods transform a multichannel signal into binaural ear signals, typically in a single processing stage. When SRIRs are used, this involves converting the SRIR into a BRIR by means of a convolutional multichannel matrix filter and creating the ear signals by convolution with a source signal.

Parametric methods are motivated by the common RIR model introduced in Section 2.1, which distinguishes between direct sound, early reflections, and late reverberation. These methods typically render the early response with high directional accuracy, using additional information such as the DoA of the direct sound and early reflections, while rendering the late reverberation with reduced directional resolution. Parametric approaches received early attention due to their ability to efficiently adapt to listener or source movement by updating the early response, as well as their potential for computationally efficient rendering of late reverberation [17, 107–110].

The next sections introduce both concepts in detail and discuss the corresponding contributions of this thesis.

4.1 Direct Rendering

Direct SRIR-based binaural rendering of a virtual source combines a source signal s, an (SH-domain or microphone-domain) SRIR $\boldsymbol{a}$ containing the room-acoustic cues, and a binaural rendering filter $\boldsymbol{w}_{\text{bin}}$, derived from a set of HRTFs $\boldsymbol{h}$ encoding the head-related cues. In the frequency domain, the resulting binaural signal for one ear is given by

$$s_{\text{bin}}(f) = \boldsymbol{w}_{\text{bin}}^{\top}(f)\, \boldsymbol{a}(f)\, s(f). \tag{4.1}$$

This equation shows that the binaural rendering can be implemented equivalently as either first converting the SRIR to a BRIR via the operation $\boldsymbol{w}_{\text{bin}}^{\top}(f)\boldsymbol{a}(f)$, and then filtering the source signal, or first convolving the source signal with the SRIR via $\boldsymbol{a}(f)s(f)$ and subsequently applying the binaural rendering filter. Since direct rendering does not exploit any specific structure of the SRIR, any linear, signal-independent binaural rendering method may be employed to obtain $\boldsymbol{w}_{\text{bin}}$.

Three main classes of methods have been proposed in the literature: Ambisonic binaural rendering, beamforming-based binaural reproduction (BFBR), and binaural signal matching (BSM). In Paper F, we proposed an end-to-end magnitude least squares (eMagLS) rendering framework that unifies and extends recent variants of these approaches.

In the following, a unified perspective on direct binaural rendering is provided, clarifying relationships between Ambisonic rendering, BFBR, BSM, and eMagLS. For notational clarity and to emphasize conceptual relations, the following formulations are presented in a simplified form, omitting regularization terms that are nonetheless essential in practical implementations.

Ambisonic Rendering Binaural rendering from Ambisonics typically assumes ideal Ambisonic signals. In the context of SRIR-based rendering, this corresponds to assuming that the SRIR $\boldsymbol{a}$ is represented in the SH domain without errors from discrete spatial sampling. In practice, an approximate SH representation is obtained for arbitrary arrays using a general ATF-based encoder, such as (2.24). The resulting SH coefficients are then subject to errors at low frequencies due to encoder regularization and at high frequencies due to SH order truncation and spatial aliasing [55, 111].

Early Ambisonic binaural renderers were inspired by loudspeaker-based reproduction. The Ambisonic signal (or SRIR) is first rendered to a set of virtual loudspeakers (VLs), followed by convolution with HRIRs corresponding to the loudspeaker directions [112, 113]. Assuming an approximately uniformly distributed loudspeaker setup that supports the SH order, the VL signals are obtained via the inverse SH transform (2.5), yielding binaural rendering filters after multiplying with the HRTFs,

$$\boldsymbol{w}_{\text{bin}}^{\text{VL}}(f) = \frac{4\pi}{Q}\boldsymbol{Y}_Q^{\top}\boldsymbol{h}(f), \tag{4.2}$$

where $\boldsymbol{Y}_Q \in \mathbb{R}^{Q\times(N+1)^2}$ contains the SHs evaluated at the Q HRTF directions.

Alternatively, the HRTFs themselves may be expressed in the SH domain [114, 115]. LS optimal rendering filters are then obtained by minimizing the squared error between the HRTFs and the SH-domain reconstruction evaluated in the same directions,

$$\begin{aligned}\boldsymbol{w}_{\text{bin}}^{\text{LS}}(f) &= \arg\min_{\boldsymbol{w}} \left\|\boldsymbol{Y}_Q\boldsymbol{w}(f) - \boldsymbol{h}(f)\right\|_2^2 \\ &= (\boldsymbol{Y}_Q^\top\boldsymbol{Y}_Q)^{-1}\boldsymbol{Y}_Q^\top\,\boldsymbol{h}(f).\end{aligned} \tag{4.3}$$

This solution can be interpreted either as the SH transform of the HRTFs via (2.6) or as the optimal filters that map ideal Ambisonic plane waves to HRTFs from the same directions. By virtue of plane-wave decomposition, these filters are applicable to any Ambisonic signal, provided that the directional sampling of plane-wave directions is sufficiently uniform over the sphere. The LS approach is equivalent to the VL approach if $\boldsymbol{Y}_Q^\top\boldsymbol{Y}_Q = \frac{Q}{4\pi}\boldsymbol{I}$, which holds due to the orthonormality of the SHs when the sampling grid supports spherical quadrature up to order N and suitable quadrature weights are employed. A spherical t-design with $t \geq 2N$ satisfies this condition and, since all quadrature weights are equal, does not require explicit weighting [41].

In practice, rendering low-order Ambisonic signals using such low-order SH representations of HRTFs leads to spatio-spectral coloration [116–118]. Motivated by the Duplex theory [119], it has been shown that the effective HRTF directionality and the resulting coloration can be reduced by removing interaural time differences at high frequencies [120]. This idea was later generalized by optimizing only the magnitude of the rendered response above a transition frequency f_c [121]. The resulting magnitude least squares (magLS) renderer combines the LS solution at low frequencies with magnitude optimization at high frequencies,

$$\boldsymbol{w}_{\text{bin}}^{\text{MLS}}(f) = \arg\min_{\boldsymbol{w}} \begin{cases} \left\|\boldsymbol{Y}_Q\boldsymbol{w}(f) - \boldsymbol{h}(f)\right\|_2^2, & f \leq f_\text{c}, \\ \left\|\,|\boldsymbol{Y}_Q\boldsymbol{w}(f)| - |\boldsymbol{h}(f)|\,\right\|_2^2, & f > f_\text{c}. \end{cases} \tag{4.4}$$

Although this non-convex problem admits convex relaxation formulations, in practice, iterative solvers that propagate phase information from neighboring frequency bins are commonly preferred [41, 122].

Beamforming-Based Binaural Reproduction Ambisonic VL rendering can be interpreted as steering SH-domain beamformers in a set of directions, followed by convolution with HRIRs. This interpretation may have motivated the development of beamforming-based rendering, first for spherical and later for arbitrary arrays [123–127]. The resulting framework was later referred to as BFBR.

In its most general form, the approach relies on the array signal model from (2.8). It designs a set beamformers $\boldsymbol{W}_\text{BF} \in \mathbb{C}^{Q\times M}$ and weights their output by the factors

$\boldsymbol{\alpha} \in \mathbb{R}^Q$ to balance a non-uniform beamformer grid. To obtain rendering filters, the weighted beamformers are multiplied by HRTFs from corresponding directions,

$$\boldsymbol{w}_{\text{bin}}^{\text{BFBR}}(f) = \boldsymbol{W}_{\text{BF}}^{\top}(f)\,\text{diag}\{\boldsymbol{\alpha}\}\,\boldsymbol{h}(f). \tag{4.5}$$

For SMAs with $M \geq (N+1)^2$ microphones and sufficiently uniform spatial coverage to resolve all SH modes, BFBR using maximum directivity index (max-DI) beamformers is equivalent to Ambisonic VL rendering[1] [127].

Binaural Signal Matching While BFBR enables flexible rendering for arbitrary arrays, it does not provide a direct optimal array-to-binaural mapping. Motivated by the concept of synthesizing array directivity patterns via a filter-and-sum operation, the virtual artificial head method directly optimizes the match between a filtered set of ATFs $\boldsymbol{P}_{\text{A}}$ and the HRTF [128, 129],

$$\begin{aligned}\boldsymbol{w}_{\text{bin}}^{\text{BSM}}(f) &= \arg\min_{\boldsymbol{w}} \left\| \boldsymbol{P}_{\text{A}}^{\top}(f)\boldsymbol{w}(f) - \boldsymbol{h}(f) \right\|_2^2 \\ &= (\boldsymbol{P}_{\text{A}}^{*}(f)\boldsymbol{P}_{\text{A}}^{\top}(f))^{-1}\boldsymbol{P}_{\text{A}}^{*}(f)\boldsymbol{h},\end{aligned} \tag{4.6}$$

where $(\cdot)^*$ denotes complex conjugation. The concept was later also derived from a signal-dependent perspective and referred to as BSM [130].

The operator $(\boldsymbol{P}_{\text{A}}^{*}\boldsymbol{P}_{\text{A}}^{\top})^{-1}\boldsymbol{P}_{\text{A}}^{*}$ can be interpreted as a bank of beamformers, allowing BSM to be cast as a BFBR variant [131]. Moreover, comparing (4.6) with the Ambisonic LS renderer (4.3) reveals that both approaches are equivalent when the microphone array constitutes an ideal Ambisonic receiver, i.e., when $\boldsymbol{P}_{\text{A}}^{\top} = \boldsymbol{Y}_Q$. Consequently, the BSM approach suffers from the same coloration issues as LS Ambisonic rendering.

End-to-End Magnitude Least Squares Rendering In Paper F, we extended the Ambisonic magLS renderer to explicitly incorporate array characteristics by replacing the ideal SH model with an array signal model, resulting in the eMagLS renderer. While initially derived for spherical and equatorial arrays in Paper F and [132], the approach generalizes naturally to arbitrary arrays using measured or simulated ATFs as used in [133, 134] and Paper B.

[1] The Ambisonic SMA encoder from (2.23) transforms signals to the SH domain and inverts the radial terms:

$$\boldsymbol{E}_{\text{SMA}}(f) = \text{diag}\{\boldsymbol{b}(f)\}^{-1}\,(\boldsymbol{Y}_{\text{M}}^{\top}\boldsymbol{Y}_{\text{M}})^{-1}\boldsymbol{Y}_{\text{M}}^{\top}.$$

VL signals in the HRTF directions are then obtained by applying the inverse SH transform. These signals are identical to those produced by a max-DI beamformer designed for the same array [127],

$$\boldsymbol{W}_{\text{max-DI}}(f) = \boldsymbol{Y}_Q\,\text{diag}\{\boldsymbol{b}(f)\}^{-1}\,(\boldsymbol{Y}_{\text{M}}^{\top}\boldsymbol{Y}_{\text{M}})^{-1}\boldsymbol{Y}_{\text{M}}^{\top} = \boldsymbol{Y}_Q\boldsymbol{E}_{\text{SMA}}(f).$$

The eMagLS rendering filters are obtained by replacing the SH matrix $\boldsymbol{Y}_Q$ in (4.4) with (analytic, measured, or simulated) ATFs $\boldsymbol{P}_\mathrm{A}$,

$$\boldsymbol{w}_\mathrm{bin}^\mathrm{EMLS}(f) = \arg\min_{\boldsymbol{w}} \begin{cases} \left\| \boldsymbol{P}_\mathrm{A}^\top(f)\boldsymbol{w}(f) - \boldsymbol{h}(f) \right\|_2^2, & f \leq f_\mathrm{c}, \\ \left\| \left| \boldsymbol{P}_\mathrm{A}^\top(f)\boldsymbol{w}(f) \right| - \left| \boldsymbol{h}(f) \right| \right\|_2^2, & f > f_\mathrm{c}. \end{cases} \tag{4.7}$$

Unlike the magLS filters, these eMagLS filters operate directly on the microphone signals rather than an SH representation, which motivated the term *end-to-end* magLS. Nevertheless, the general ATF-based signal model facilitates another possibility: the design of an Ambisonic binaural renderer that is aware of array and encoder properties. Consider an arbitrary Ambisonic encoder $\boldsymbol{E}$ that transforms the ATFs $\boldsymbol{P}_\mathrm{A}$ to the SH domain,

$$\boldsymbol{P}_N(f) = \boldsymbol{E}(f)\boldsymbol{P}_\mathrm{A}(f). \tag{4.8}$$

The corresponding Ambisonic eMagLS renderer is then

$$\boldsymbol{w}_\mathrm{bin}^\mathrm{A-EMLS}(f) = \arg\min_{\boldsymbol{w}} \begin{cases} \left\| \boldsymbol{P}_N^\top(f)\boldsymbol{w}(f) - \boldsymbol{h}(f) \right\|_2^2, & f \leq f_\mathrm{c}, \\ \left\| \left| \boldsymbol{P}_N^\top(f)\boldsymbol{w}(f) \right| - \left| \boldsymbol{h}(f) \right| \right\|_2^2, & f > f_\mathrm{c}. \end{cases} \tag{4.9}$$

In Paper F, the Ambisonic variant (4.9) is termed *eMagLS*, while the ATF-based formulation (4.7) is referred to as *eMagLS2*. The optimization for both eMagLS variants is solved analogously to magLS. In practice, both require regularization.

Comparing (4.7) with the BSM formulation (4.6) reveals that the low-frequency LS stage of eMagLS is equivalent to signal-independent BSM, up to regularization. Consequently, the recently proposed BSM-magLS method [131] is equivalent to eMagLS. Similar equivalences hold for recent array-aware Ambisonic methods [135], where Ambisonic signal matching (ASM) corresponds to the LS encoder (2.24) [56], and array-aware magLS binaural rendering coincides with Ambisonic eMagLS (4.9).

By explicitly accounting for array and HRTF characteristics within a single optimization, eMagLS mitigates errors that arise at both low and high frequencies. At low frequencies, higher-order spatial modes of the array are weak, which limits spatial resolution and necessitates regularization. At high frequencies, the spatial complexity of the sound field exceeds the spatial resolution provided by the finite number of microphones, resulting in spatial aliasing. While aliasing cannot be fully eliminated, the joint optimization ensures that the interaction between array response and HRTFs is balanced in a least-squares sense, yielding a diffuse-field equalized rendering whose spatially averaged energy response matches the target. In two independent studies, eMagLS performed best among signal-independent binaural renderers [133, 134].

Figure 4.1 compares the binaural renderers for a plane wave impinging from 50° azimuth onto the eight-microphone smartglasses array used in Paper B. Ambisonic encoding with maximum order $N = 1$ is performed with a diffuse-field equalized variant of the LS encoder defined in (2.24). The BFBR method employs a set of

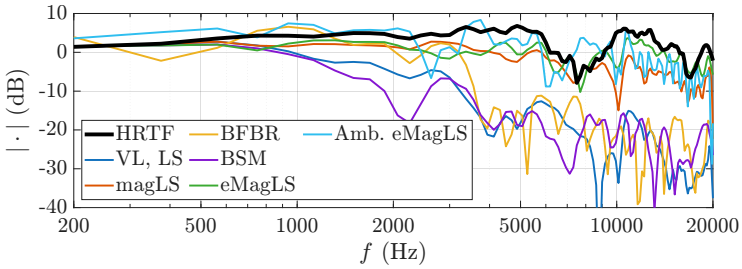


Figure 4.1: Comparison of direct binaural renderers for a single plane wave. While all renderers yield a similar magnitude response at low frequencies, the magLS-based approaches achieve a more accurate reproduction of the reference HRTF at mid and high frequencies, and the eMagLS renderers are most accurate.

MVDR beamformers. The renderers use the HRTF set from [136], for which the VL and LS renderers are equivalent. While all renderers perform similarly at low frequencies, the magLS-based renderers achieve a more accurate reproduction of the corresponding HRTF at mid and high frequencies, and the array-aware eMagLS variants provide the most accurate match.

4.2 Parametric Rendering

While direct rendering is robust in that it does not rely on additional estimation or modeling assumptions, its directional accuracy is fundamentally limited by the number and spatial distribution of microphones. Parametric rendering methods aim to overcome this limitation by estimating additional acoustic parameters, allowing the direct sound and early reflections to be rendered with increased directional accuracy. The late response, by contrast, is typically rendered using computationally efficient techniques with lower directional resolution to reduce overall rendering cost. While parametric renderers often target loudspeaker-based reproduction, they can generally be adapted for binaural playback, in the simplest case by using the VL approach that was introduced for Ambisonic direct rendering. In many cases, also dedicated binaural variants have been proposed.

Early parametric approaches rendered acoustic environments by geometrically simulating a small number of early reflections and convolving them with corresponding HRIRs [107–110]. These methods were later extended to include diffuse late reverberation, typically synthesized using FDNs [17, 18].

Subsequent methods targeted SRIRs in a stricter sense, and are typically designed for measured responses rather than a small set of explicitly modeled early reflections. As illustrated in Figure 4.2, these approaches typically contain separate analysis,

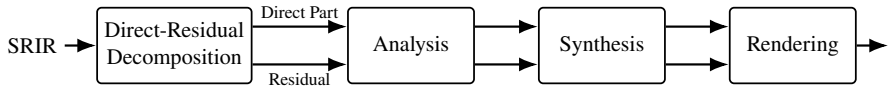


Figure 4.2: Parametric SRIR processing commonly includes a decomposition into direct and residual components and subsequent individual analysis, synthesis, and rendering of both components.

synthesis, and rendering stages for the direct part (direct sound and prominent reflections) and the residual reverberation. After the SRIR is decomposed into direct and residual components, the analysis stage for the direct part typically identifies the direct sound and salient reflections and estimates their times of arrival, DoAs, and spectra. The residual is commonly analyzed in terms of the RT, DRR, energy decay curve, frequency-dependent energy, and diffuseness. The synthesis stage creates direct sound and reflections with increased directional resolution based on the determined spectra and DoAs. The late reverberation is either decorrelated, resynthesized using shaped decaying noise, or reproduced via artificial reverberators. Finally, the rendering stage maps the resulting signals to loudspeaker or binaural outputs and combines them with the source signal.

One of the earliest methods in this category is spatial impulse response rendering (SIRR), which decomposes first-order Ambisonic SRIRs into a directional component, rendered using single-direction amplitude panning based on direction estimates derived from the pseudo-intensity vector, and a diffuse component, which is decorrelated prior to rendering [137]. In contrast, the spatial decomposition method (SDM) assumes an open microphone array and employs time-difference-of-arrival-based direction estimates at each time sample, assigning a rendering direction without explicitly separating the response into directional and diffuse streams [138]. SIRR was later extended to higher-order Ambisonic SRIRs by applying the same principles within multiple directional sectors, yielding separate directional and diffuse streams per sector [139]. Alternative SDM formulations based on Ambisonic SRIRs have also been proposed [140], along with modifications specifically targeting binaural reproduction [10]. More recent works largely follow the general processing structure illustrated in Figure 4.2, but differ in their assumptions about the microphone array as well as in the specific algorithms used for direct-residual separation, detection, direction, and spectral estimation of early reflections, and residual rendering [141–144].

Several studies have focused on improving individual components of this framework. In particular, accurate rendering of the late part of SRIRs with correct binaural coherence has been investigated in [145, 146], while efficient methods for late reverberation rendering have been studied in [147–149]. In a similar spirit, this thesis contributes detailed investigations and novel methods for a central component of the parametric rendering framework: the separation of direct and residual SRIR components.

Most of the previously mentioned parametric SRIR renderers employ some kind of direct and residual separation, often without a detailed analysis of their effectiveness. For example, SIRR separates the energy in time-frequency bins into direct and diffuse streams by using a diffuseness estimate based on the length of the pseudo-intensity vector [137]. Later works typically employ a peak detection algorithm for direct sound and early reflections and extract them using a set of beamformers [141–144]. While these methods showed the individual success of their full processing, the effectiveness of extracting reflections in realistic scenarios was not investigated in detail.

That motivated two works that proposed algorithms for the separation of direct sound and early reflections from measured SRIRs in Paper G and Paper H. Specifically, the works provide methods that aim to extract direct sound and strong early reflections while preserving the directional properties of the reflections and the residual that is left after the direct components have been directionally extracted.

Direct and Residual Separation using a Plane-Wave Model The first work in Paper G specifically investigates beamformer design for SMAs, operating in the SH domain. Like the eMagLS renderer that uses a comprehensive array signal model instead of assuming ideal Ambisonic receivers, the proposed beamformer exploits the comprehensive array signal model to improve the extraction of reflections from an SRIR. In a nutshell, the method describes reflections as plane waves that are picked up by an SMA and encoded to the SH domain similar to (4.8),

$$\boldsymbol{p}_N(f,\Omega) = \boldsymbol{E}(f)\boldsymbol{p}_\mathrm{A}(f,\Omega)s(f)\,. \tag{4.10}$$

Here, s is the spectrum of the plane wave, $\boldsymbol{E}$ is an arbitrary Ambisonic encoder, and $\boldsymbol{p}_\mathrm{A}$ is an ATF for incidence direction Ω following the analytic model from (2.20) and (2.21). After a reflection is detected and its DoA estimated, the method isolates the reflection with a temporal window and extracts the plane-wave spectrum s with a matched-filter beamformer based on the ATF model. The model from (4.10) then serves as reflection model so that the reflection can be removed from the SRIR via frequency-domain subtraction to obtain the SRIR residual

$$\boldsymbol{r}_N(f) = \boldsymbol{a}_N(f) - \hat{\boldsymbol{p}}_N(f), \tag{4.11}$$

where $\boldsymbol{a}_N$ is the isolated section of the Ambisonic SRIR containing the reflection and $\hat{\boldsymbol{p}}_N$ is the estimated Ambisonic plane wave based on the estimated reflection spectrum and direction. For rendering, the reflection is synthesized with arbitrary directional resolution as a plane wave using its estimated spectrum and DoA, thereby avoiding distortion in the extracted reflection $\hat{\boldsymbol{p}}_N$ due to microphone array scattering and Ambisonic encoding.

While Paper G showed the effectiveness of this method for extracting individual strong early reflections and the method generalizes to arbitrary arrays using the general signal model from (2.18), it relies on several other estimators whose performance of reliably detecting reflections and their arrival directions likely degrades in complex scenarios.

Direct and Residual Separation via Subspace Decomposition The method proposed in PAPER H addresses these shortcomings by avoiding the need for separate reflection detection and DoA estimation algorithms. Instead of relying on temporal windowing and an explicit estimate of how many reflections arrive from which directions, it works directly with an estimate of the residual and separates direct and residual components using a subspace-based approach. Inspired by the subspace decomposition concept that has been applied in various signal processing algorithms [90, 150–153], it exploits the underlying assumption that the captured SRIR spans a high-dimensional space, of which the direct components only occupy a lower-dimensional subspace. As shown in PAPER H, this assumption holds for the early part of SRIRs when a sufficient number of microphones is available. Since the method does not depend on an explicit microphone array model, it can be applied in the same way to SRIRs represented in either the microphone or the Ambisonic domain.

An overview of the procedure is given in Figure 4.3. The method sequentially processes the SRIR in blocks from the end toward the beginning, assuming that the tail of the response contains no individual prominent reflections but only consists of the residual. Thus, the initial SRIR block serves as the initial residual estimate. For each subsequent block, the generalized singular value decomposition (GSVD) of the current SRIR block $\boldsymbol{X}$ and the current residual estimate $\boldsymbol{N}$ is then computed as

$$\boldsymbol{X} = \boldsymbol{V}_\mathrm{x}\boldsymbol{\Sigma}_\mathrm{x}\boldsymbol{\Phi}^\top , \tag{4.12}$$

$$\boldsymbol{N} = \boldsymbol{V}_\mathrm{n}\boldsymbol{\Sigma}_\mathrm{n}\boldsymbol{\Phi}^\top , \tag{4.13}$$

where $\boldsymbol{V}_\mathrm{x}$ and $\boldsymbol{V}_\mathrm{n}$ contain left singular vectors, $\boldsymbol{\Phi}$ are their common right singular vectors, and the diagonal matrices $\boldsymbol{\Sigma}_\mathrm{x}$ and $\boldsymbol{\Sigma}_\mathrm{n}$ contain the corresponding singular values.

If the sum of the generalized singular values (GSVs) exceeds a threshold based on the time-averaged GSV sum of the residual, indicating an energetic peak in the current SRIR block that exceeds the residual energy, a subspace decomposition is performed. The SRIR block $\boldsymbol{X}$ is then split into a direct component $\boldsymbol{X}_\mathrm{s}$ and a residual component $\boldsymbol{X}_\mathrm{n}$ by constructing $\boldsymbol{X}_\mathrm{s}$ from the Q_s largest singular values and their associated singular vectors, while the remaining components are assigned to the residual,

$$\boldsymbol{X}_\mathrm{s} = \boldsymbol{V}_\mathrm{x}\boldsymbol{\Sigma}_\mathrm{x}\boldsymbol{\Gamma}_\mathrm{s}\boldsymbol{\Phi}^\top , \tag{4.14}$$

$$\boldsymbol{X}_\mathrm{n} = \boldsymbol{V}_\mathrm{x}\boldsymbol{\Sigma}_\mathrm{x}\boldsymbol{\Gamma}_\mathrm{n}\boldsymbol{\Phi}^\top . \tag{4.15}$$

Here, $\boldsymbol{\Gamma}_\mathrm{s}$ and $\boldsymbol{\Gamma}_\mathrm{n}$ are binary diagonal matrices that select the corresponding components.

By iterating this procedure from the end of the SRIR toward its beginning while continuously updating the residual estimate, direct sound and prominent early reflections can be separated from the residual without the need for additional parameter estimation. An example of an SRIR decomposed into direct and residual components using this approach is shown in Figure 4.4.

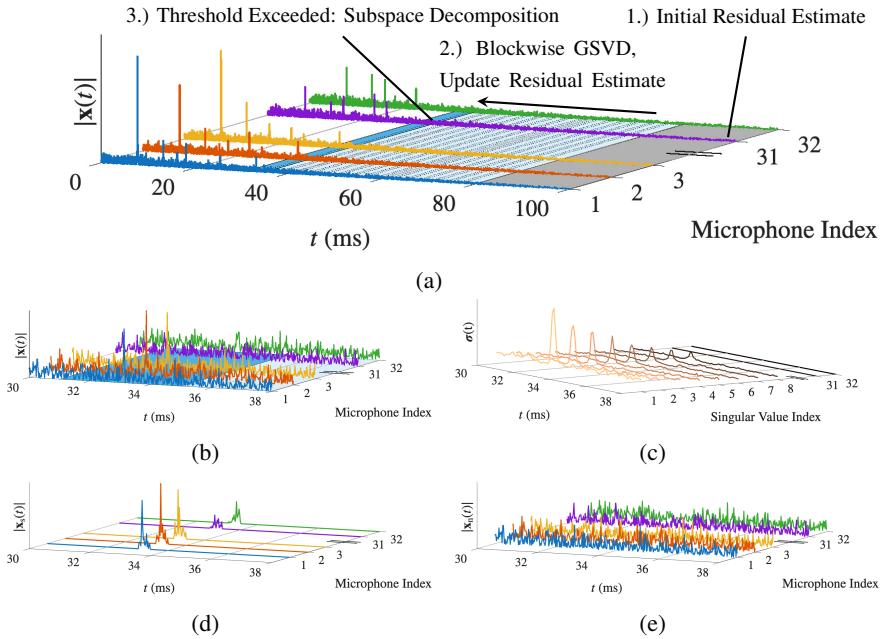


Figure 4.3: Direct and residual subspace decomposition of an SRIR. In (a), the algorithm initializes a residual estimate from the tail of the SRIR and then processes the response block by block toward its beginning, computing a GSVD for each block. Blocks for which the sum of the GSVs remains below a detection threshold are treated as residual, while blocks exceeding the threshold are decomposed into direct and residual components. In (b), a zoomed-in segment of the SRIR contains a prominent reflection. In (c), this reflection is clearly visible as a peak in the largest GSVs, whereas the smallest GSVs do not exhibit a pronounced peak. As a result, the reflection is assigned to the direct component in (d), while the remaining residual in (e) does not contain this reflection. Figure adapted from Paper H.

Discussion Subspace-based separation of direct and residual components provides an alternative to array-model-based approaches and avoids several practical limitations associated with parametric reflection modeling. By operating directly on the structure of the measured SRIR and exploiting differences in subspace dimensionality, this approach remains effective at higher frequencies, where the aliased response is highly sensitive to small changes in source direction, making reliable modeling impractical in real-world scenarios, and in scenarios where multiple reflections fall within the same processing block and cannot be cleanly separated using parametric techniques.

From a theoretical standpoint, the method assumes that the direct part occupies a lower-dimensional subspace, which requires a sufficient number of microphones to

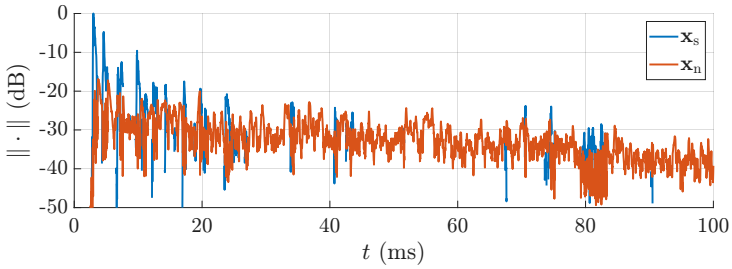


Figure 4.4: Subspace decomposition of an SRIR into its direct part $\boldsymbol{x}_\mathrm{s}$, containing the direct sound and prominent reflections, and its residual $\boldsymbol{x}_\mathrm{n}$. Figure adapted from Paper H.

yield a singular direct-part spatial covariance matrix in the presence of a plane wave. In practice, however, this condition does not need to be strictly fulfilled for the method to remain useful. Even when the separation between direct and residual subspaces is imperfect, perceptually meaningful improvements can still be obtained. This has been shown, for instance, by its use with SRIRs from an array of only four microphones in an SDM-based renderer, where the subspace decomposition method improved rendering performance [154].

The method also proved useful in the extrapolation of SRIRs to support 6DoF rendering from only a small number of SRIR measurements [16]. In that context, improved separation of direct and residual components enabled perceptually meaningful extensions of the sound field beyond the region covered by the original measurements.

Chapter 5

Conclusions and Future Work

This thesis addressed the problems of estimating and rendering real-world acoustics for plausible binaural auralization, with a particular focus on applications in VR/AR. The central objective was to enable spatially accurate, perceptually convincing reproduction of room acoustics using microphone arrays that are compact, wearable, and suitable for practical deployment.

To this end, the thesis made contributions at both ends of the virtual acoustics pipeline: the estimation of SRIRs from microphone array recordings, and the binaural rendering of these responses for headphone playback. A unifying theme throughout the work is the explicit separation of room-dependent and array-dependent components, enabling array-independent representations that facilitate flexible processing and rendering.

In the context of SRIR estimation, a modular framework based on pseudo-reference signal estimation was introduced and systematically investigated. By reformulating blind multichannel system identification as an informed estimation problem, the proposed approach overcomes fundamental limitations of classical cross-relation-based methods when applied to long acoustic impulse responses. The framework integrates dereverberation, beamforming, and LS system identification, and was shown to provide accurate SRIR estimates from naturally occurring sounds such as speech. Extensive evaluations demonstrated that the resulting estimates preserve perceptually relevant room acoustic parameters, including reverberation time and direct-to-reverberant energy ratio, even in the presence of noise and moderate DoA errors.

The framework was further extended to scenarios involving moving microphone arrays, which are characteristic of wearable devices. By incorporating array rotation and translation into the signal model, the thesis demonstrated that motion can be compensated for and, in some cases, exploited to improve spatial resolution. Using

circular-harmonic-based representations, it was shown that combining multiple observations from different array poses enables the reconstruction of sound-field coefficients with increased spatial bandwidth, allowing accurate SRIR estimation beyond the native spatial aliasing limits of the array. Experimental results with head-worn microphone arrays under natural head motion revealed practical limitations of the method, particularly in spatial accuracy and at high frequencies. Nevertheless, omnidirectional RIR estimates obtained under these conditions still facilitated perceptually promising rendering and are of practical benefit.

On the rendering side, the thesis contributed a unified perspective on binaural rendering methods. By formulating Ambisonic rendering, beamforming-based binaural reproduction, and binaural signal matching within a common optimization framework, fundamental equivalences between existing approaches were clarified. Building on this analysis, the eMagLS renderer was introduced. By explicitly accounting for array characteristics and encoding errors, eMagLS mitigates spatio-spectral coloration caused by limited spatial resolution, order truncation, and spatial aliasing. The framework generalizes naturally to arbitrary microphone arrays and unifies several recently proposed array-aware binaural rendering methods.

Finally, the thesis investigated parametric SRIR rendering with an emphasis on the separation of direct components (direct sound and early reflections) and residual reverberation. Two complementary methods were introduced and analyzed. The first is based on an explicit array signal model, enabling the extraction and resynthesis of direct sound and early reflections while preserving the spatial characteristics of the residual reverberation. The second method approaches the problem from a signal-subspace perspective, decomposing the SRIR into direct and residual components without relying on explicit reflection detection or detailed array-specific modeling assumptions. Owing to its increased robustness and reduced sensitivity to modeling errors, this subspace-based approach is often preferable in practical scenarios. Together, the two methods establish a foundation for improving parametric SRIR-based rendering approaches.

Future Work While this thesis focused on model-based signal processing methods for SRIR estimation and binaural rendering, recent advances in machine learning have demonstrated impressive performance on related problems, including RIR estimation and acoustic matching. Rather than replacing physical model-based approaches, an important direction for future work lies in combining such models with data-driven methods.

Model-based formulations, such as those developed in this thesis, provide strong inductive biases that learned components can exploit to mitigate limitations from simplifying assumptions needed for analytical tractability. Hybrid approaches have therefore attracted increasing attention in both the signal processing and machine learning communities, where physics-informed and model-aware neural networks

have been shown to improve generalization, reduce training data requirements, and accelerate convergence [155–160].

In acoustic and audio signal processing, common assumptions include Gaussian signal models, linear and time-invariant system behavior, and independence across time frames and frequency bins. While these assumptions enable tractable formulations, they often lead to ill-posed or underdetermined inverse problems in realistic scenarios involving noise, reverberation, and multiple sources. Learning-based components can address these limitations by capturing complex correlations and nonlinear relationships across space, frequency, and time, effectively introducing data-driven priors on physically plausible sound-field evolution, for example under array motion.

Hybrid audio signal processing thus enables the design of efficient, high-performing algorithms by keeping physically motivated assumptions that promote robustness and low complexity, while introducing data-driven components where model limitations affect performance most. Recent work by the author explored two such hybrid approaches, improving neural-network-based Ambisonic encoding from microphone arrays by learning to enhance the residual after model-based encoding [161], and proposing a lightweight multiple-input multiple-output speech enhancement method that incorporates a learned spatial covariance matrix into a model-based enhancement algorithm [162]. Within modular processing frameworks such as the SRIR estimation approach proposed in this thesis, learning-based modules can be integrated to replace only those components that limit overall performance, while preserving the underlying model-based structure for system identification. For example, the model-based dereverberation stage could be replaced with a neural module to better handle mixture signals and dynamic acoustic conditions. In the future, such hybrid approaches will enable robust and efficient algorithms that are crucial for practical implementations, but further research is needed to determine which model assumptions are beneficial and which learned components are necessary.

Chapter 6

Summary of the Appended Publications

Paper A: Blind Estimation of Spatial Room Impulse Responses Using a Pseudo Reference Signal

Paper A introduces a modular framework for blind SRIR estimation that reformulates blind multichannel system identification as an informed problem. Instead of directly estimating acoustic impulse responses via channel cross-relations, which is shown to fail for realistic RIR lengths, a pseudo-reference signal is first estimated from the microphone array signals using dereverberation and beamforming. This reference enables the use of standard informed system identification techniques to estimate SRIRs. The method is evaluated using circular, spherical, and equatorial microphone arrays, demonstrating accurate reconstruction of early reflections and energy decay, as well as reliable RT and DRR estimation across array geometries.

Paper B: Blind Identification of Binaural Room Impulse Responses From Smart Glasses

Paper B evaluates the estimation framework from Paper A using a head-worn microphone array integrated into smartglasses. The work demonstrates that BRIRs can be reliably estimated under stationary conditions using speech as the excitation signal. A comprehensive statistical analysis shows that the estimated responses capture perceptually relevant room parameters such as RT and DRR with high accuracy and robustness against noise, interfering sources, and moderate direction-of-arrival errors. In addition to objective evaluation, a perceptual experiment assesses the plausibility of binaural renderings based on the estimated BRIRs. The results indicate that the estimated responses enable perceptually convincing acoustic matching and are rated as more similar to the real room than renderings based on responses measured in

different rooms of comparable size.

Paper C: Spatial Room Impulse Response Identification from Rotating Equatorial Microphone Arrays

Paper C extends SRIR estimation to rotating equatorial microphone arrays using a CH-domain signal model. The paper shows that array rotation can be exploited to increase directional resolution beyond the limitations of a static array. Different rotation speeds are analyzed in both simulations and measurements to study their influence on estimation accuracy and convergence behavior. The results demonstrate that the estimated SRIRs achieve an accuracy comparable to measured SRIRs from static arrays with significantly more microphones. These findings show that array rotation can actively enhance spatial resolution rather than merely introducing additional modeling challenges.

Paper D: Spatial Room Impulse Response Estimation From a Moving Microphone Array

Paper D generalizes the rotation-aware estimation framework to translational motion. By expressing the sound field using CH coefficients that are valid over a spatial region, SRIR estimation becomes possible from moving arrays. The study shows that linear translational motion can be exploited to improve spatial resolution and obtain reliable estimates beyond the array's nominal spatial aliasing frequency. At higher frequencies, however, estimation accuracy is ultimately limited by the finite order of the CH representation and by the translation distance.

Paper E: Identification and Matching of Room Acoustics With Moving Head-Worn Microphone Arrays

Paper E evaluates the complete estimation framework, including dereverberation and beamforming, in realistic scenarios with moving head-worn microphone arrays and natural head motion. The study investigates whether motion-aware SRIR estimation enables perceptually convincing acoustic matching in practical AR-like settings. While reliable RT and DRR estimation remains possible, directional rendering based on the estimated SRIR does not consistently outperform an alternative omnidirectional approach, which nevertheless achieves perceptually convincing results.

Paper F: End-to-End Magnitude Least Squares Binaural Rendering of Spherical Microphone Array Signals

Paper F focuses on direct binaural rendering from spherical and equatorial microphone array signals. It proposes an array-aware eMagLS renderer that explicitly accounts for array characteristics during binaural reproduction. The approach reduces spatio-spectral coloration and provides diffuse-field-equalized rendering at high frequencies. The work contributes to a unified understanding of array-dependent and array-independent rendering strategies and lays the foundation for general array-aware, magnitude-optimal binaural rendering.

Paper G: Spatial Subtraction of Reflections from Room Impulse Responses Measured with a Spherical Microphone Array

Paper G presents a method for separating reflections from SRIRs measured with SMAs. Using an SH-based array signal model tailored to SMAs, individual reflection components are spatially modeled and subtracted from the measured responses. This facilitates improved extraction of direct sound and early reflections while preserving the directional characteristics of the residual, which is particularly relevant for parametric rendering approaches that process direct and residual components separately.

Paper H: Direct and Residual Subspace Decomposition of Spatial Room Impulse Responses

Paper H proposes a subspace-based method for decomposing SRIRs into direct and residual components. Instead of relying on an explicit array model, the method employs a GSVD to separate direct sound and prominent reflections from the residual component based on a residual estimate derived from the SRIR. The approach demonstrates improved robustness at higher frequencies, where spatial aliasing limits the model-based method introduced in Paper G. The subspace decomposition approach facilitates parametric rendering methods that enhance directional sharpness while preserving the spatial characteristics of late reverberation.

References

[1] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, "Auralization - An Overview", *J. Audio Eng. Soc.* 41(11), pp. 861–875, 1993.

[2] D. Satongar, C. Pike, Y. W. Lam, and A. I. Tew, "The influence of headphones on the localization of external loudspeaker sources", *J. Audio Eng. Soc.* 63(10), pp. 799–810, 2015.

[3] C. Schneiderwind, A. Neidhardt, and D. Meyer, "Comparing the effect of different open headphone models on the perception of a real sound source", *Proc. 150th Conv. Audio Eng. Soc.* pp. 1–10, 2021.

[4] A. Mülleder, P. Lladó, and N. Meyer-Kahlen, "Comparison of Passive Transparent Headphones for Augmented Reality Audio", *Proc. AES Int. Conf. on Headphone Technology*, pp. 1–8, 2025.

[5] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the authenticity of individual dynamic binaural synthesis", *J. Acoust. Soc. Am.* 142(4), pp. 1784–1795, 2017.

[6] A. Lindau and S. Weinzierl, "Assessing the plausibility of virtual acoustic environments", *Acta Acustica united with Acustica* 98(5), pp. 804–810, 2012.

[7] S. A. Wirler, N. Meyer-Kahlen, and S. J. Schlecht, "Towards transfer-plausibility for evaluating mixed reality audio in complex scenes", *Proc. AES Int. Conf. on Audio for Virtual and Augmented Reality*, pp. 1–10, 2020.

[8] N. Meyer-Kahlen, S. J. Schlecht, S. Amengual Garí, and T. Lokki, "Testing Auditory Illusions in Augmented Reality: Plausibility, Transfer-Plausibility, and Authenticity", *J. Audio Eng. Soc.* 72(11), pp. 797–812, 2024.

[9] N. Meyer-Kahlen. "Transfer-Plausible Acoustics for Augmented Reality". PhD thesis. Aalto University, 2024.

[10] S. V. Amengual Garí, J. M. Arend, P. T. Calamia, and P. W. Robinson, "Optimizations of the spatial decomposition method for binaural reproduction", *J. Audio Eng. Soc.* 68(12), pp. 959–976, 2020.

[11] D. Fantini, G. Presti, M. Geronazzo, R. Bona, A. G. Privitera, and F. Avanzini, "Co-immersion in Audio Augmented Virtuality: The Case Study of a Static and Approximated Late Reverberation Algorithm", *IEEE Transactions on Visualization and Computer Graphics* 29(11), pp. 4472–4482, 2023.

[12] H. Himmelein, T. Lübeck, D. Bau, and C. Pörschmann, "Evaluating the Plausibility of Binaural Ambisonics and Parametric Renderings", *J. Audio Eng. Soc.* 73(11), pp. 722–733, 2025.

[13] K. Müller and F. Zotter, "Auralization based on multi-perspective ambisonic room impulse responses", *Acta Acustica* 6(25), pp. 1–18, 2020.

[14] O. Puomio, T. Pihlajakuja, and T. Lokki, "Sound rendering with early reflections extracted from a measured spatial room impulse response", *Proc. Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–8, 2021.

[15] T. McKenzie, N. Meyer-Kahlen, R. Daugintis, L. McCormack, S. J. Schlecht, and V. Pulkki, "Perceptually informed interpolation and rendering of spatial room impulse responses for room transitions", *Proc. 24th Int. Congress on Acoustics*, pp. 1–11, 2022.

[16] T. Deppisch, S. V. Amengual Garí, P. Calamia, and J. Ahrens, "Perceptual Evaluation of Spatial Room Impulse Response Extrapolation by Direct and Residual Subspace Decomposition", *Proc. AES Int. Conf. on Audio for Virtual and Augmented Reality*, pp. 1–10, 2022.

[17] J.-M. Jot, V. Larcher, and O. Warusfel, "Digital Signal Processing Issue in the Context of Binaural and Transaural Stereophony", *Proc. 98th Conv. Audio Eng. Soc.* 1995.

[18] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating Interactive Virtual Acoustic Environments", *J. Audio Eng. Soc.* 47(9), pp. 675–705, 1999.

[19] A. Benoit, A. Politis, S. J. Schlecht, and V. Välimäki, "Directional Feedback Delay Network", *J. Audio Eng. Soc.* 67(10), pp. 752–762, 2019.

[20] A. Neidhardt, C. Schneiderwind, and F. Klein, "Perceptual Matching of Room Acoustics for Auditory Augmented Reality in Small Rooms - Literature Review and Theoretical Framework", *Trends in Hearing* 26, pp. 1–22, 2022.

[21] N. Meyer-Kahlen, S. V. Garí, I. Ananthabhotla, and P. Calamia, "A Two-Dimensional Threshold Test for Reverberation Time and Direct-to-Reverberant Ratio", *Proc. Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–8, 2023.

[22] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of Room Acoustic Parameters: The ACE Challenge", *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24(10), pp. 1681–1693, 2016.

[23] H. Gamper and I. J. Tashev, "Blind reverberation time estimation using a convolutional neural network", *Proc. IEEE Int. Workshop on Acoustic Signal Enhancement*, pp. 136–140, 2018.

[24] W. Mack, S. Deng, and E. A. Habets, "Single-channel blind direct-to-reverberation ratio estimation using masking", *Proc. INTERSPEECH*, pp. 5066–5070, 2020.

[25] P. Srivastava, A. Deleforge, and E. Vincent, "Blind Room Parameter Estimation Using Multiple-Multichannel Speech Recordings", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–5, 2021.

[26] P. Götz, C. Tuna, A. Walther, and E. A. P. Habets, "Online reverberation time and clarity estimation in dynamic acoustic conditions", *J. Acoust. Soc. Am.* 153(6), pp. 3532–3542, 2023.

[27] S. Saini and J. Peissig, "Blind Room Acoustic Parameters Estimation Using Mobile Audio Transformer", *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–5, 2023.

[28] C. Ick, A. Mehrabi, and W. Jin, "Blind Acoustic Room Parameter Estimation Using Phase Features", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

[29] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics", *J. Acoust. Soc. Am.* 65(4), pp. 943–950, 1979.

[30] J. A. Moorer, "About This Reverberation Bisiness", *Computer Music Journal* 3(2), pp. 13–28, 1985.

[31] J.-D. Polack, "Modifying chambers to play billiards: The foundations of reverberation theory", *Acta Acustica united with Acustica* 76(6), pp. 256–272, 1992.

[32] J. D. Polack, "Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics", *Applied Acoustics* 38(2-4), pp. 235–244, 1993.

[33] J.-M. Jot, "An analysis/synthesis approach to real-time artificial reverberation", *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 221–224, 1992.

[34] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space", *Proc. of the National Academy of Sciences* 113(48), pp. E7856–E7865, 2016.

[35] B. Rafaely. *Fundamentals of Spherical Array Processing*. 2nd. Springer, 2019.

[36] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A round robin on room acoustical simulation and auralization", *J. Acoust. Soc. Am.* 145(4), pp. 2746–2760, 2019.

[37] E. G. Williams. *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. Academic Press, 1999.

[38] H. Teutsch. *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*. Springer, 2007.

[39] J. Ahrens. *Analytic Methods of Sound Field Synthesis*. Springer Berlin, Heidelberg, 2012.

[40] N. Ueno and S. Koyama, "Sound Field Estimation: Theories and Applications", *Foundations and Trends in Signal Processing* 19(1), pp. 1–98, 2025.

[41] F. Zotter and M. Frank. *Ambisonics, A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer, 2019.

[42] M. A. Gerzon, "Periphohy: With-Height Sound Reproduction." *J. Audio Eng. Soc.* 21(1), pp. 2–10, 1973.

[43] M. A. Gerzon, "Ambisonics in Multichannel Broadcasting and Video", *J. Audio Eng. Soc.* 33(11), pp. 859–871, 1985.

[44] M. Poletti, "A Unified Theory of Horizontal Holographic Sound Systems", *J. Audio Eng. Soc.* 48(12), pp. 1155–1182, 2000.

[45] S. Gannot, D. Burshtein, and E. Weinstein, "Signal Enhancement Using Beamforming and Nonstationarity with Applications to Speech", *IEEE Transactions on Signal Processing* 49(8), pp. 1614–1626, 2001.

[46] I. Cohen, "Relative Transfer Function Identification Using Speech Signals", *IEEE Transactions on Speech and Audio Processing* 12(5), pp. 451–459, 2004.

[47] R. Talmon, I. Cohen, and S. Gannot, "Relative Transfer Function Identification Using Convolutive Transfer Function Approximation", *IEEE Transactions on Audio, Speech and Language Processing* 17(4), pp. 546–555, 2009.

[48] B. Laufer, R. Talmon, and S. Gannot, "Relative Transfer Function Modeling for Supervised Source Localization", *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, 2013.

[49] S. Doclo, W. Kellermann, S. Makino, and S. Nordholm, "Multichannel Algorithms for Assisted Signal Enhancement Listening Devices: Exploiting spatial diversity using multiple microphones", *IEEE Signal Processing Magazine* 32(2), pp. 18–30, 2015.

[50] D. Fejgin and S. Doclo, "Coherence-Based Frequency Subset Selection for Binaural RTF-vector-based Direction of Arrival Estimation for Multiple Speakers", *Proc. Int. Workshop on Acoustic Signal Enhancement*, pp. 1–5, 2022.

[51] H. L. Van Trees. *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. New York: John Wiley & Sons, 2002.

[52] A. Bastine, L. Birnie, T. D. Abhayapala, P. Samarasinghe, and V. Tourbabin, "Ambisonics Capture using Microphones on Head-worn Device of Arbitrary Geometry", *Proc. European Signal Processing Conference*, pp. 309–313, 2022.

[53] J. Meyer and G. W. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield", *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. II-1781-II–1784, 2002.

[54] B. Rafaely, "Spatial sampling and beamforming for spherical microphone arrays", *Proc. Workshop on Hands-free Speech Communication and Microphone Arrays*, pp. 5–8, 2008.

[55] A. Politis and H. Gamper, "Comparing modeled and measurement-based spherical harmonic encoding filters for spherical microphone arrays", *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 224–228, 2017.

[56] S. Moreau, J. Daniel, and S. Bertet, "3D Sound Field Recording with Higher Order Ambisonics – Objective Measurements and Validation of Spherical Microphone", *Proc. 120th Conv. Audio Eng. Soc.* pp. 1–24, 2006.

[57] M. A. Gerzon, "The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound", *J. Audio Eng. Soc.* 23, pp. 402–404, 1975.

[58] C. Schörkhuber and R. Höldrich, "Ambisonic Microphone Encoding with Covariance Constraint", *Proc. Int. Conf. on Spatial Audio*, 2017.

[59] L. Mccormack, A. Politis, R. Gonzalez, T. Lokki, and V. Pulkki, "Parametric Ambisonic Encoding of Arbitrary Microphone Arrays", *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30, pp. 2062–2075, 2022.

[60] K. Abed-Meraim, W. Qiu, I. Member, and Y. Hua, "Blind system identification", *Proceedings of the IEEE* 85(8), pp. 1310–1322, 1997.

[61] M. I. Gürelli and C. L. Nikias, "EVAM: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals", *IEEE Transactions on Signal Processing* 43(1), pp. 134–149, 1995.

[62] G. Xu, H. Liu, L. Tong, and T. Kailath, "A Least-Squares Approach to Blind Channel Identification", *IEEE Transactions on Signal Processing* 43(12), pp. 2982–2993, 1995.

[63] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification", *IEEE Transactions on Signal Processing* 51(1), pp. 11–24, 2003.

[64] S. Gannot and M. Moonen, "Subspace Methods for Multimicrophone Speech Dereverberation", *EURASIP J. on Applied Signal Processing* 2003(11), pp. 1074–1090, 2003.

[65] M. A. Haque and M. K. Hasan, "Noise Robust Multichannel Frequency-Domain LMS Algorithms for Blind Channel Identification", *IEEE Signal Processing Letters* 15, pp. 305–308, 2008.

[66] B. Jo and P. Calamia, "Robust blind multichannel identification based on a phase constraint and different lp-norm constraints", *Proc. 28th European Signal Processing Conference*, pp. 1966–1970, 2021.

[67] Z. Liao, F. Xiong, J. Luo, M. Cai, E. S. Chng, J. Feng, and X. Zhong, "Blind Estimation of Room Impulse Response from Monaural Reverberant Speech with Segmental Generative Neural Network", *Proc. INTERSPEECH*, pp. 2723–2727, 2023.

[68] A. Ratnarajah, I. Ananthabhotla, V. K. Ithapu, P. Hoffmann, D. Manocha, and P. Calamia, "Towards Improved Room Impulse Response Estimation for Speech Recognition", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

[69] S. Lee, H. S. Choi, and K. Lee, "Yet Another Generative Model for Room Impulse Response Estimation", *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–5, 2023.

[70] C. J. Steinmetz, V. K. Ithapu, and P. Calamia, "Filtered Noise Shaping for Time Domain Room Impulse Response Estimation from Reverberant Speech", *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 221–225, 2021.

[71] L. Lalay, M. Fontaine, and R. Badeau, "Unified Variational and Physics-aware Model for Room Impulse Response Estimation", *Proc. INTERSPEECH*, pp. 3818–3822, 2025.

[72] K. Lee, J. Seo, K. Choi, S. Lee, and B. S. Chon, "Room Impulse Response Estimation in a Multiple Source Environment", *Proc. AES Int. Conf. on Spatial and Immersive Audio*, pp. 1–11, 2023.

[73] F. Lluís and N. Meyer-Kahlen, "Blind Spatial Impulse Response Generation from Separate Room- and Scene-Specific Information", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1–5, 2024.

[74] A. Ratnarajah, S. Ghosh, S. Kumar, P. Chiniya, and D. Manocha, "AV-RIR: Audio-Visual Room Impulse Response Estimation", *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 27154–27165, 2024.

[75] E. Moliner, J. M. Lemercier, S. Welker, T. Gerkmann, and V. Valimaki, "BUDDy: Single-Channel Blind Unsupervised Dereverberation with Diffusion Models", *Proc. 18th Int. Workshop on Acoustic Signal Enhancement*, pp. 120–124, 2024.

[76] J.-M. Lemercier, E. Moliner, S. Welker, V. Välimäki, and T. Gerkmann, "Unsupervised Blind Joint Dereverberation and Room Acoustics Estimation with Diffusion Models", *IEEE Transactions on Audio, Speech and Language Processing* 33, pp. 2244–2258, 2025.

[77] S. Saini, I. Engel, and J. Peissig, "An end-to-end approach for blindly rendering a virtual sound source in an audio augmented reality environment", *EURASIP J. on Audio, Speech, and Music Processing* 2024(16), pp. 1–24, 2024.

[78] P. Götz, G. D. Santo, S. J. Schlecht, V. Välimäki, and E. A. P. Habets, "Matching Reverberant Speech Through Learned Acoustic Embeddings and Feedback Delay Networks", *arXiv:2510.23158*, pp. 1–5, 2025.

[79] H. Kim, L. Hernaggi, P. J. Jackson, and A. Hilton, "Immersive spatial audio reproduction for VR/AR using room acoustic modelling from 360° images", *Proc. 26th IEEE Conf. on Virtual Reality and 3D User Interfaces*, pp. 120–126, 2019.

[80] L. Remaggi, H. Kim, A. Neidhardt, A. Hilton, and P. J. Jackson, "Perceived quality and spatial impression of room reverberation in VR reproduction from measured images and acoustics", *Proc. Int. Congress on Acoustics*, pp. 3361–3368, 2019.

[81] H. Kon and H. Koike, "An auditory scaling method for reverb synthesis from a single two-dimensional image", *Acoustical Science and Technology* 41(4), pp. 675–685, 2020.

[82] N. Singh, J. Mentch, J. Ng, M. Beveridge, and I. Drori, "Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis", *Proc. IEEE Int. Conf. on Computer Vision*, pp. 286–295, 2021.

[83] C. Chen, R. Gao, P. Calamia, and K. Grauman, "Visual Acoustic Matching", *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 18836–18846, 2022.

[84] A. Somayazulu, C. Chen, and K. Grauman, "Self-Supervised Visual Acoustic Matching", *Proc. 37th Conference on Neural Information Processing Systems*, pp. 1–19, 2023.

[85] A. Perez-Lopez, A. Politis, and E. Gomez, "Blind reverberation time estimation from ambisonic recordings", *Proc. IEEE 22nd Int. Workshop on Multimedia Signal Processing*, pp. 1–6, 2020.

[86] N. Meyer-Kahlen and S. J. Schlecht, "Blind Directional Room Impulse Response Parameterization from Relative Transfer Functions", *Proc. IEEE Int. Workshop on Acoustic Signal Enhancement*, pp. 1–5, 2022.

[87] S. Haykin. *Adaptive Filter Theory*. Fifth Edit. Pearson, 2015.

[88] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening", *IEEE Transactions on Audio, Speech and Language Processing* 20(10), pp. 2707–2720, 2012.

[89] T. Nakatani, N. Kamo, M. Delcroix, and S. Araki, "Multi-Stream Diffusion Model for Probabilistic Integration of Model-Based and Data-Driven Speech Enhancement", *Proc. 18th Int. Workshop on Acoustic Signal Enhancement*, pp. 65–69, 2024.

[90] R. O. Schmidt, "Multiple emitter location and parameter estimation", *IEEE Transactions on Antennas and Propagation* 34(3), pp. 276–280, 1986.

[91] N. Meyer-Kahlen and T. Deppisch, "Direct-to-Reverberant Energy Ratio Estimation and Extrapolation from Own Speech", *Proc. 33rd European Signal Processing Conference*, pp. 321–325, 2025.

[92] T. Ajdler, L. Sbaiz, and M. Vetterli, "Dynamic measurement of room impulse responses using a moving microphone", *J. Acoust. Soc. Am.* 122(3), pp. 1636–1645, 2007.

[93] N. Hahn and S. Spors, "Continuous measurement of spatial room impulse responses using a non-uniformly moving microphone", *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 205–208, 2017.

[94] F. Katzberg, R. Mazur, M. Maass, P. Koch, and A. Mertins, "Sound-field measurement with moving microphones", *J. Acoust. Soc. Am.* 141(5), pp. 3220–3235, 2017.

[95] F. Katzberg, R. Mazur, M. Maass, P. Koch, and A. Mertins, "A Compressed Sensing Framework for Dynamic Sound-Field Measurements", *IEEE/ACM Transactions on Audio Speech and Language Processing* 26(11), pp. 1962–1975, 2018.

[96] F. Katzberg, M. Maass, and A. Mertins, "Spherical Harmonic Representation For Dynamic Sound-Field Measurements", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 426–430, 2021.

[97] K. MacWilliam, T. Dietzen, R. Ali, and T. van Waterschoot, "State-space estimation of spatially dynamic room impulse responses using a room acoustic model-based prior", *Frontiers in Signal Processing* 4, pp. 1–18, 2024.

[98] J. Brunnström, M. B. Møller, and M. Moonen, "Bayesian sound field estimation using moving microphones", *IEEE Open Journal of Signal Processing*, pp. 1–10, 2025.

[99] N. A. Gumerov and R. Duraiswami. *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions*. Amsterdam: Elsevier Science, 2004.

[100] P. Samarasinghe, T. Abhayapala, and M. Poletti, "Wavefield analysis over large areas using distributed higher order microphones", *IEEE Transactions on Audio, Speech and Language Processing* 22(3), pp. 647–658, 2014.

[101] J. G. Tylka and E. Y. Choueiri, "Soundfield Navigation using an Array of Higher-Order Ambisonics Microphones", *Proc. AES Conf. on Audio for Virtual and Augmented Reality*, pp. 1–10, 2016.

[102] N. Ueno, S. Koyama, and H. Saruwatari, "Sound field recording using distributed microphones based on harmonic analysis of infinite order", *IEEE Signal Processing Letters* 25(1), pp. 135–139, 2018.

[103] F. Schultz and S. Spors, "Data-based binaural synthesis including rotational and translatory head-movements", *Proc. 52nd AES Int. Conf.* pp. 1–11, 2013.

[104] N. Hahn and S. Spors, "Modal Bandwidth Reduction in Data-based Binaural Synthesis including Translatory Head-movements", *Proc. Annual German Conference on Acoustics (DAGA)*, pp. 1–4, 2015.

[105] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers", *IEEE Transactions on Speech and Audio Processing* 9(6), pp. 697–707, 2001.

[106] H. Møller, "Fundamentals of binaural technology", *Applied Acoustics* 36(3-4), pp. 171–218, 1992.

[107] S. H. Foster, E. M. Wenzel, and R. M. Taylor, "Real Time Synthesis of Complex Acoustic Environments", *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–2, 1991.

[108] D. R. Begault, "Binaural Auralization and Perceptual Veridicality", *Proc. Proc. 93rd Conv. Audio Eng. Soc.* pp. 1–23, 1992.

[109] H. Lehnert and J. Blauert, "Principles of binaural room simulation", *Applied Acoustics* 36(3-4), pp. 259–291, 1992.

[110] J. P. Vian and J. Martin, "Binaural room acoustics simulation: Practical uses and applications", *Applied Acoustics* 36(3-4), pp. 293–305, 1992.

[111] S. Moreau, J. Daniel, and S. Bertet, "3D Sound Field Recording with Higher Order Ambisonics - Objective Measurements and Validation of a 4th Order Spherical Microphone", *Proc. 120th Conv. Audio Eng. Soc.* pp. 1–24, 2006.

[112] J.-M. Jot, V. Larcher, and J.-M. Pernaux, "A Comparative Study of 3-D Audio Encoding and Rendering Techniques", *Proc. AES 16th Int. Conference*, pp. 281–300, 1999.

[113] M. Noisternig, A. Sontacchi, T. Musil, and R. Höldrich, "A 3D Ambisonic Based Binaural Sound Reproduction System", *Proc. AES 24th Int. Conf. on Multichannel Audio*, pp. 1–5, 2003.

[114] J. Daniel. “Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia”. PhD thesis. Université de Paris, 2001.

[115] D. Menzies, “W-Panning and O-Format, Tools for Object Spatialization”, *Proc. Int. Conf on Auditory Display*, pp. 1–8, 2002.

[116] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely, “Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution”, *J. Acoust. Soc. Am.* 133(5), pp. 2711–2721, 2013.

[117] B. Bernschütz, A. Vázquez Giner, C. Pörschmann, and J. M. Arend, “Binaural reproduction of plane waves with reduced modal order”, *Acta Acustica united with Acustica* 100(5), pp. 972–983, 2014.

[118] J. Sheaffer and B. Rafaely, “Equalization strategies for binaural room impulse response rendering using spherical arrays”, *Proc. IEEE 28th Conv. of Electrical and Electronics Engineers in Israel*, pp. 1–5, 2014.

[119] L. Rayleigh, “On our perception of sound direction”, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13(74), pp. 214–232, 1907.

[120] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, “Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint”, *J. Acoust. Soc. Am.* 143(6), pp. 3616–3627, 2018.

[121] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, “Binaural Rendering of Ambisonic Signals via Magnitude Least Squares”, *Proc. Annual German Conference on Acoustics (DAGA)*, pp. 339–342, 2018.

[122] C. Hold, N. Meyer-Kahlen, and V. Pulkki, “Magnitude-Least-Squares Binaural Ambisonic Rendering with Phase Continuation”, *Proc. Annual German Conference on Acoustics (DAGA)*, pp. 6–9, 2023.

[123] R. Duraiswami, D. N. Zotkin, Z. Li, E. Grassi, N. A. Gumerov, and L. S. Davis, “High Order Spatial Audio Capture and its Binaural Head-Tracked Playback over Headphones with HRTF Cues”, *Proc. 119th Conv. Audio Eng. Soc.* pp. 1–16, 2005.

[124] A. M. O’Donovan, D. N. Zotkin, and R. Duraiswami, “Spherical Microphone Array Based Immersive Audio Scene Rendering”, *Proc. 14th Int. Conf. on Auditory Display*, pp. 1–8, 2008.

[125] S. Spors, H. Wierstorf, and M. Geier, “Comparison of modal versus delay-and-sum beamforming in the context of data-based binaural synthesis”, *Proc. 132nd Conv. Audio Eng. Soc.* pp. 822–837, 2012.

[126] N. R. Shabtai, “Optimization of the directivity in binaural sound reproduction beamforming”, *J. Acoust. Soc. Am.* 138(5), pp. 3118–3128, 2015.

[127] I. Ifergan and B. Rafaely, “On the selection of the number of beamformers in beamforming-based binaural reproduction”, *EURASIP J. on Audio, Speech, and Music Processing* 2022(6), pp. 1–17, 2022.

[128] E. Rasumow, M. Blau, M. Hansen, S. Doclo, S. Van der Par, V. Mellert, and D. Püschel, "Robustness of virtual artificial head topologies with respect to microphone positioning", *Proc. Forum Acusticum*, pp. 2251–2256, 2011.

[129] E. Rasumow, M. Blau, S. Doclo, S. Van der Par, M. Hansen, D. Puschel, and V. Mellert, "Perceptual Evaluation of Individualized Binaural Reproduction Using a Virtual Artificial Head", *J. Audio Eng. Soc.* 65(6), pp. 448–459, 2017.

[130] L. Madmoni, J. Donley, V. Tourbabin, and B. Rafaely, "Beamforming-based Binaural Reproduction by Matching of Binaural Signals", *Proc. AES Int. Conf. on Audio for Virtual and Augmented Reality*, pp. 1–8, 2020.

[131] L. Madmoni, Z. Ben-Hur, J. Donley, V. Tourbabin, and B. Rafaely, "Design and analysis of binaural signal matching with arbitrary microphone arrays and listener head rotations", *EURASIP J. on Audio, Speech, and Music Processing* 2025(11), pp. 1–20, 2025.

[132] H. Helmholz, T. Deppisch, and J. Ahrens, "End-to-End Magnitude Least Squares Binaural Rendering for Equatorial Microphone Arrays", *Proc. Annual German Conference on Acoustics (DAGA)*, pp. 1679–1682, 2023.

[133] B. Stahl and S. Riedel, "Perceptual Comparison of Dynamic Binaural Reproduction Methods for Head-Mounted Microphone Arrays", *Proc. 155th Conv. Audio Eng. Soc.* pp. 1–10, 2023.

[134] L. McCormack, N. Meyer-Kahlen, D. L. Alon, Z. Ben-Hur, S. V. Amengual Gari, and P. Robinson, "Six-Degrees-of-Freedom Binaural Reproduction of Head-Worn Microphone Array Capture", *J. Audio Eng. Soc.* 71(10), pp. 638–649, 2023.

[135] Y. Gayer, V. Tourbabin, Z. Ben-Hur, D. L. Alon, and B. Rafaely, "Array-Aware Ambisonics and HRTF Encoding for Binaural Reproduction With Wearable Arrays", *arXiv:2507.11091*, pp. 1–11, 2025.

[136] B. Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100", *Proc. Annual German Conference on Acoustics (DAGA)*, pp. 592–595, 2013.

[137] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: Analysis and synthesis", *J. Audio Eng. Soc.* 53(12), pp. 1115–1127, 2005.

[138] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses", *J. Audio Eng. Soc.* 61(1/2), pp. 17–28, 2013.

[139] L. McCormack, V. Pulkki, A. Politis, O. Scheuregger, and M. Marschall, "Higher-Order Spatial Impulse Response Rendering: Investigating the Perceived Effects of Spherical Order, Dedicated Diffuse Rendering, and Frequency Resolution", *J. Audio Eng. Soc.* 68(5), pp. 338–354, 2020.

[140] M. Zaunschirm, M. Frank, and F. Zotter, "BRIR synthesis using first-order microphone arrays", *Proc. 144th Conv. Audio Eng. Soc.* pp. 1–10, 2018.

[141] P. Coleman, A. Franck, P. J. Jackson, R. J. Hughes, L. Remaggi, and F. Melchior, "Object-based reverberation for spatial audio", *J. Audio Eng. Soc.* 65(1-2), pp. 66–77, 2017.

[142] P. Stade, J. Arend, and C. Pörschmann, "A parametric model for the synthesis of binaural room impulse responses", *Proc. Meetings on Acoustics*, pp. 1–12, 2017.

[143] L. McCormack, N. Meyer-Kahlen, and A. Politis, "Spatial Reconstruction-Based Rendering of Microphone Array Room Impulse Responses", *J. Audio Eng. Soc.* 71(5), pp. 267–280, 2023.

[144] H. Sun, H. Y. Zhu, M. T. D. Nguyen, V. Nguyen, C.-t. Lin, and C. T. Jin, "From RIR to BRIR: A Sparse Recovery Beamforming Approach for Virtual Binaural Sound Rendering", *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1231–1235, 2024.

[145] F. Menzer and C. Faller, "Obtaining binaural room impulse responses from B-format impulse responses", *Proc. 125th Conv. Audio Eng. Soc.* pp. 912–919, 2008.

[146] F. Menzer, C. Faller, and H. Lissek, "Obtaining binaural room impulse responses from b-format impulse responses using frequency-dependent coherence matching", *IEEE Transactions on Audio, Speech and Language Processing* 19(2), pp. 396–405, 2011.

[147] N. Agus, H. Anderson, J. M. Chen, S. Lui, and D. Herremans, "Minimally simple binaural room modeling using a single feedback delay network", *J. Audio Eng. Soc.* 66(10), pp. 791–807, 2018.

[148] C. Kirsch, J. Poppitz, T. Wendt, S. Van De Par, and S. D. Ewert, "Computationally Efficient Spatial Rendering of Late Reverberation in Virtual Acoustic Environments", *Proc. Immersive and 3D Audio: From Architecture to Automotive*, pp. 1–8, 2021.

[149] J. Fagerström, N. Meyer-Kahlen, S. J. Schlecht, and V. Välimäki, "Binaural Dark-Velvet-Noise Reverberator", *Proc. 27th Int. Conf. on Digital Audio Effects*, pp. 246–253, 2024.

[150] D. W. Tufts, R. Kumaresan, and I. Kirsteins, "Data Adaptive Signal Estimation By Singular Value Decomposition of a Data Matrix", *Proceedings of the IEEE* 70(6), pp. 684–685, 1982.

[151] R. Roy and T. Kailath, "ESPRIT - Estimation of Signal Parameters Via Rotational Invariance Techniques", *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37(7), pp. 984–995, 1989.

[152] Y. Ephraim and H. L. Van Trees, "A Signal Subspace Approach for Speech Enhancement", *IEEE Transactions on Speech and Audio Processing* 3(4), 1995.

[153] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement", *IEEE Transactions on Signal Processing* 50(9), pp. 2230–2244, 2002.

[154] L. Gölles and M. Frank, "Ambisonic Spatial Decomposition Method with salient / diffuse separation", *Proc. 158th Conv. Audio Eng. Soc*, pp. 1–9, 2025.

[155] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems

involving nonlinear partial differential equations", *Journal of Computational Physics* 378, pp. 686–707, 2019.

[156] P. J. Baddoo, B. Herrmann, B. J. McKeon, J. N. Kutz, and S. L. Brunton, "Physics-informed dynamic mode decomposition", *Proc. Royal Society A* 479(2771), pp. 1–22, 2023.

[157] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-Based Deep Learning", *Proceedings of the IEEE* 111(5), pp. 465–499, 2023.

[158] A. Bonfanti, R. Santana, M. Ellero, and B. Gholami, "On the generalization of PINNs outside the training domain and the hyperparameters influencing it", *Neural Computing and Applications* 36, pp. 22677–22696, 2024.

[159] R. A. McCarthy, Y. Zhang, S. A. Verburg, W. F. Jenkins, and P. Gerstoft, "Machine Learning in Acoustics: A Review and Open-source Repository", *npj Acoustics* 1(18), pp. 1–18, 2025.

[160] C. Meng, S. Griesemer, D. Cao, S. Seo, and Y. Liu, "When physics meets machine learning: a survey of physics-informed machine learning", *Machine Learning for Computational Science and Engineering* 1(20), pp. 1–23, 2025.

[161] T. Deppisch, Y. Gao, M. Mittal, B. Stahl, C. Hold, D. Alon, and Z. Ben-Hur, "Residual Learning for Neural Ambisonics Encoders", *arXiv:2601.18322*, pp. 1–6, 2026.

[162] T. Deppisch, "Direction-Preserving MIMO Speech Enhancement Using a Neural Covariance Estimator", *arXiv:2604.11179*, pp. 1–6, 2026.