Department of Information and Communications Engineering

Transfer-Plausible Acoustics for Augmented Reality

Nils Meyer-Kahlen

Aalto University

DOCTORAL THESES

Aalto University publication series
DOCTORAL THESES 139/2024

Transfer-Plausible Acoustics for Augmented Reality

Nils Meyer-Kahlen

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall F239a of the school on 26 July 2024 at 12 noon.

Aalto University
School of Electrical Engineering
Department of Information and Communications Engineering
Virtual Acoustics Group

Supervising professor
Prof. Tapio Lokki, Aalto University, Finland

Thesis advisors
Prof. Sebastian J. Schlecht, Friedrich-Alexander Universität Erlangen-Nürnberg, Germany
Dr. Philip Robinson, Reality Labs Research, USA

Preliminary examiners
Dr. Fabian Brinkmann, Technische Universität Berlin, Germany
Dr. Chris Pike, Sonos, Inc, UK

Opponent
Prof. Christoph Pörschmann, Technische Hochschule Köln, Germany

Aalto University publication series
DOCTORAL THESES 139/2024

ISBN 978-952-64-1912-1 (printed)
ISBN 978-952-64-1913-8 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-1913-8

Unigrafia Oy
Helsinki 2024

Finland

Printed matter
4041-0619

Author
Nils Meyer-Kahlen

Name of the doctoral thesis
Transfer-Plausible Acoustics for Augmented Reality

Publisher School of Electrical Engineering

Unit Department of Information and Communications Engineering

Series Aalto University publication series DOCTORAL THESES 139/2024

Field of research Acoustics and Audio Signal Processing

Manuscript submitted 11 April 2024 **Date of the defence** 26 July 2024

Permission for public defence granted (date) 17 June 2024 **Language** English

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

Abstract

Augmented reality (AR) telepresence systems aim to present visual and auditory "holograms" of conversation partners via head-mounted displays and transparent headphones. These systems require binaural audio that adapts not only to the user's orientation and position but also to their acoustic environment. Many fundamental technologies for such real-time, binaural auralization systems have been developed over the years. These virtual acoustic systems were often tested in direct comparison to a high-quality reference rendering, so the implied objective for the system's development was often indistinguishability from a reference. However, differences were usually audible in such tests, at least for non-ideal, practically relevant systems. When developing future AR systems, two questions arise: "Why exactly do such discrepancies occur?" and "What are meaningful objectives and evaluation paradigms other than indistinguishability from a reference?"
First, finding reasons for discrepancies involves a detailed understanding of specific rendering methods, underlying models, and their violations. Two fundamental properties of a parametric spatial room impulse response processing technique are studied as examples.
Second, as an objective that leads to meaningful AR evaluation paradigms, one option is to assess if auditory illusions are evoked, i.e., whether a listener believes a virtual sound source to be real. This work introduces the transfer-plausibility paradigm, which evaluates if a virtual source creates an auditory illusion, even in the presence of other, real sound sources.
In summary, Publication I and Publication II discuss fundamental properties of spatial room impulse response processing techniques: Publication I shows how direction-of-arrival estimation based on the pseudo intensity vector depends on anisotropy in the late reverberation. Publication II investigates how perceptual roughness can occur in spatial room impulse response rendering based on broadband directional assignment.
Publication III and Publication IV deal with problems more closely related to AR. Publication III proposes an approach for blind spatial room impulse response estimation using a pseudo-reference signal. Publication IV demonstrates auditory modeling-based quantification of impairments caused by so-called transparent headphones used for AR.
Publication V and Publication VI introduce the notion of transfer-plausibility and compare it against other paradigms. The results suggest that even non-ideal virtual acoustic renderings are comparable in transfer-plausibility tests. Publication VII presents an experiment about the inability for self-localization using position-dependent room acoustic differences. The thesis concludes by presenting opportunities for future transfer-plausibility tests and a proposed model for describing differences in experimental paradigms by their sensitivity to auditory similarity, context, and artifacts.

Keywords Virtual Acoustics, Augmented Reality, Spatial Room Impulse Response Processing, Room Acoustics, Plausibility

ISBN (printed) 978-952-64-1912-1 **ISBN (pdf)** 978-952-64-1913-8

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2024

Pages 181 **urn** http://urn.fi/URN:ISBN:978-952-64-1913-8

Abstract

Augmented reality (AR) telepresence systems aim to present visual and auditory "holograms" of conversation partners via head-mounted displays and transparent headphones. These systems require binaural audio that adapts not only to the user's orientation and position but also to their acoustic environment.

Many fundamental technologies for such real-time, binaural auralization systems have been developed over the years. These virtual acoustic systems were often tested in direct comparison to a high-quality reference rendering, so the implied objective for the system's development was often indistinguishability from a reference. However, differences were usually audible in such tests, at least for non-ideal, practically relevant systems. When developing future AR systems, two questions arise: "Why exactly do such discrepancies occur?" and "What are meaningful objectives and evaluation paradigms other than indistinguishability from a reference?"

First, finding reasons for discrepancies involves a detailed understanding of specific rendering methods, underlying models, and their violations. Two fundamental properties of a parametric spatial room impulse response processing technique are studied as examples.

Second, as an objective that leads to meaningful AR evaluation paradigms, one option is to assess if auditory illusions are evoked, i.e., whether a listener believes a virtual sound source to be real. This work introduces the transfer-plausibility paradigm, which evaluates if a virtual source creates an auditory illusion, even in the presence of other, real sound sources.

In summary, Publication I and Publication II discuss fundamental properties of spatial room impulse response processing techniques: Publication I shows how direction-of-arrival estimation based on the pseudo intensity vector depends on anisotropy in the late reverberation. Publication II investigates how perceptual roughness can occur in spatial room impulse response rendering based on broadband directional assignment.

Publication III and Publication IV deal with problems more closely related to AR. Publication III proposes an approach for blind spatial room impulse response estimation using a pseudo-reference signal. Publication IV demonstrates auditory modelling-based quantification of impairments caused by so-called transparent headphones used for AR.

Publication V and Publication VI introduce the notion of transfer-plausibility and compare it against other paradigms. The results suggest that even non-ideal virtual acoustic renderings are comparable in transfer-plausibility tests. Publication VII presents an experiment about the inability for self-localization using position-dependent room acoustic differences. The thesis concludes by presenting opportunities for future transfer-plausibility tests and a proposed model for describing differences in experimental paradigms by their sensitivity to auditory similarity, context, and artifacts.

Preface

This thesis marks an important milestone in my academic path through acoustics and audio. When I started as a Bachelor's student in Graz in 2013, I still thought I would become a recording engineer. Soon after, I was drawn to the Institute of Electronic Music and Acoustics, and all the exciting scientific questions, the knowledge, and the joy everyone got from experimenting. I am very grateful for having received the opportunity to continue my path at the Aalto Acoustics Laboratory – the best place for working on acoustics and audio that I can think of. I was especially lucky to be part of the European training network "VRACE" at the same time, where I met many wonderful colleagues and friends. Moreover, I would like to thank Meta's Reality Labs Research, where I was an intern during the time of this thesis. My gratitude also goes to the Nokia Foundation and the HPY Research Foundation for their contributions to my research. Personally, I would like to thank:

– Tapio Lokki, for making all of this possible and always having my back
– Sebastian J. Schlecht, for being a great advisor, a friend, and a source of great ideas and critical reflection about acoustics, signal processing, research, and life as a whole
– Sebastià Amengual Garí, for listening to me talk for too long more than once, but still asking the right questions in between
– Christoph Pörschmann, for acting as my opponent
– Fabian Brinkmann, and Chris Pike for being my pre-examiners
– Philip Robinson, for kicking off the project with us
– Aleksi Öyry, for being the best enabler of experimental research possible
– Ville Pulkki, for teaching me about creativity, buckets and more
– Vesa Välimäki, for teaching me to write clearly and effectively
– Leo McCormack, for demonstrating how making things that actually work is worthwhile
– Janani Fernandez, for teaching me about language, fruit, greek mythology, and many, many other things

– Georg Götz, for being a highly reliable source of important information that otherwise would have been missed
– Pedro Lladó, for in-depth discussions and clearly flagged, strong opinions whenever they were needed
– Christoph Hold, for holding the group together through reliable Thursday meetings
– Stefan Wirler, for allowing me to join in on the first steps towards transfer-plausibility that he started
– Thomas McKenzie, for looking after everyone and for the paddle boat song
– Raimundo Gonzalez, for discussions about quantum physics, obscure art, and everything in between
– Ricardo Falcón Pérez, for having such opposing but well-thought-out views about the world
– Kyung Yun Lee, Jana Nolze, and Jon Fagerström for letting me partake in their research
– Francesc Lluís, for teaching me so much about deep learning
– Winfried Lachenmayr, and Otavio Colella Gomes for taking me along on their measurement tour
– Alexander Mülleder, for the MushRoom headphones
– Aaron Geldert, Anja Hofmann, Sergio de las Heras, and Ruijie Wang, for being absolutely fantastic master students I got to work with
– everyone I crossed paths with at the lab, such as Antti Kuusinen, Vasileios Bountourakis, Alec Wright, Taeho Kim, Karolina Prawda, Madalina Natasa, Petteri Hyvärinen, Michael McCrea, Lauros Pajunen, Rapolas Daugintis, Henna Tahvanainen, Archontis Politis, Julie Meyer, Otto Puomio, Janne Riionheimo, Arif Yürek, Leonardo Fierro, Benoit Alary, Etienne Thuillier, Juho Liski, Juhani Paasonen, Eloi Moliner, Gloria Dal Santo, Lauri Savioja, Gian Marco de Bortoli, Simon Schwär, Antonio Figueroa Durán, and many more
– everyone from the VRACE network
– everyone from the Audio Team at Reality Labs Research
– Olarin Panimo

Apart from all these people from the field, I want to sincerely thank my parents and, of course, my wife, Agnes Kloft for their unconditional support.

Helsinki, June 26, 2024,

Nils Meyer-Kahlen

Contents

List of Publications

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

I Meyer-Kahlen, N., Schlecht, S.J. Directional distribution of the pseudo intensity vector in anisotropic late reverberation. *The Journal of the Acoustical Society of America*, 155(2), 1515–1526, February 2024.

II Meyer-Kahlen, N., Schlecht, S.J., Lokki, T. Perceptual roughness of spatially assigned sparse noise for rendering reverberation. *The Journal of the Acoustical Society of America*, 150(5), 3521–3531, November 2021.

III Meyer-Kahlen, N., Schlecht, S.J. Blind directional room impulse response parameterization from relative transfer functions. *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Bamberg, Germany, September 2022.

IV Lladó, P., McKenzie, T., Meyer-Kahlen, N., Schlecht, S.J. Predicting perceptual transparency of head-worn devices. *The Journal of the Audio Engineering Society*, 70 (7), 585–600, July/August 2022.

V Wirler, S., Meyer-Kahlen, N, Schlecht, S.J. Towards transfer-plausibility for evaluating mixed reality audio in complex scenes. In *AES International Conference on Audio for Virtual and Augmented Reality*, Remote, August 2020.

VI Meyer-Kahlen, N, Schlecht, S.J., Amengual Garí, S., , Lokki, T. Testing auditory illusions in augmented reality: plausibility, transfer-plausibility and authenticity. *The Journal of the Audio Engineering Society*, Submitted, 2024 .

VII Meyer-Kahlen, N., Schlecht, S.J., Lokki, T. Clearly audible differences rarely reveal where you are in a room. *The Journal of the Acoustical Society of America*, 152 (2), 877–887, August 2022.

Author's Contribution

Publication I: "Directional distribution of the pseudo intensity vector in anisotropic late reverberation"

The author had the idea and did the derivations, measurements, analyses, and writing. SJS helped with derivations and editing.

Publication II: "Perceptual roughness of spatially assigned sparse noise for rendering reverberation"

The author conceived the idea together with the other authors, ran the experiments, conducted the analyses, and wrote the paper. SJS and TL helped with editing.

Publication III: "Blind directional room impulse response parameterization from relative transfer functions"

The author had the idea, conducted the measurements and analyses, and wrote the paper. SJS helped with editing.

Publication IV: "Predicting perceptual transparency of head-worn devices"

The author conceptualized the paper together with the other authors. PL conducted the localization test, and TMK ran the coloration test, with the help of the other authors. The author conducted measurements, ran statistical analyses, and wrote the first section with the help of the other authors.

Publication V: "Towards transfer-plausibility for evaluating mixed reality audio in complex scenes"

The author helped to conceptualize the TP framework and wrote the introduction with the help of the other authors. SW implemented, ran, and analyzed the listening experiment with the help of the other authors.

Publication VI: "Testing auditory illusions in augmented reality: plausibility, transfer-plausibility and authenticity"

The ideas were discussed by all authors. The author made final decisions, implemented and ran the test, performed the measurements and the analyses, and wrote the paper. SJS, SAG, and TL helped to edit the paper.

Publication VII: "Clearly audible differences rarely reveal where you are in a room"

The idea was conceived by SJS and the author. The locoscope test was implemented with the help of Janne Pietilä, who was advised by the author and SJS. The author ran the test, performed the measurements and the analyses, and wrote the paper. SJS and TL helped to edit the paper.

List of Figures

List of Tables

Abbreviations

AAR Audio Augmented Reality

AFC Alternative Forced Choice

AR Augmented Reality

ATF Array Transfer Function

BRIR Binaural Room Impulse Response

DoF Degrees of Freedom

DRIR Directional Room Impulse Response

DRR Direct-to-reverberant energy ratio

DSP Digital Signal Processing

FDN Feedback Delay Network

HATS Head and Torso Simulator

HMD Head-mounted display

HO-SIRR Higher-Order Spatial Impulse Response Rendering

HRIR Head-Related Impulse Response

HRTF Head-Related Transfer Function

ILD Interaural Level Difference

ITD Interaural Time Difference

PIV Pseudo Intensity Vector

RIR Room Impulse Response

RT Reverberation Time

SDM Spatial Decomposition Method

SH Spherical Harmonics

SIRR Spatial Impulse Response Rendering

SMA Spherical Microphone Array

SRIR Spatial Room Impulse Response

TDoA Time Difference of Arrival

TP Transfer-Plausibility

Symbols

$\boldsymbol{A}(\boldsymbol{\Theta})$ set of ATFs, measured at directions $\boldsymbol{\Theta}$

$\boldsymbol{b}$ a binaural room impulse response

$\boldsymbol{C}$ set of microphone array response to binaural response conversion filters

$\boldsymbol{d}$ vector of beamforming coefficients

$\boldsymbol{E}$ set of microphone array encoding filters

h impulse response

$\check{\boldsymbol{h}}$ SH domain impulse response

$\boldsymbol{H}(\boldsymbol{\Theta})$ set of HRTFs, measured at directions $\boldsymbol{\Theta}$

$\boldsymbol{p}$ microphone array impulse response

$\check{\boldsymbol{v}}$ relative transfer-function

$\check{\boldsymbol{x}}$ SH domain signal

$Y_n^m(\boldsymbol{\theta})$ Spherical Harmonics (SH) of order n and degree m evaluated at $\boldsymbol{\theta}$

$\boldsymbol{Y}_N(\boldsymbol{\Theta})$ Matrix of SHs evaluated at directions $\boldsymbol{\Theta}$ up to maximal order N.

$\boldsymbol{\theta}$ direction vector

$\boldsymbol{\Theta}$ set of direction vectors

$\hat{\boldsymbol{\theta}}$ estimated direction vector

$*$ linear convolution

$.^{\mathrm{T}}$ Transposition

$.^{\mathrm{H}}$ Hermitian transposition

$.^{\dagger}$ Moore-Penrose pseudo-inverse

1. Introduction: Transfer-Plausible Virtual Acoustics

The primary motivation for studying transfer-plausible acoustics is the vision of augmented reality (AR) telepresence, in which interlocutors are shown visually via augmented reality glasses and their voices are rendered dynamically in real-time, adapted to the room acoustics of the user's environment. The concept is demonstrated in Fig. 1.1.

Undeniably, such a system takes inspiration from the science fiction world, in which blue hologram calls have foreshadowed such a technology several decades ago. Although there are still various problems to solve, functioning AR telepresence has become a much more realistic scenario during the time span of this thesis, with major technology companies developing AR technology.

This thesis contributes insights into the acoustic rendering technologies required for AR telepresence systems, such as spatial room impulse response analysis, rendering and estimation methods, and transparent headphones. Moreover, it discusses how AR systems can be evaluated.

Figure 1.1. Concept of an AR telepresence meeting. Two participants are in a room, while a third remote-end participant is rendered using head-mounted displays (HMDs) and headphones.

It could be argued that many of the basic virtual acoustic rendering technologies required for AR telepresence systems have been available for many years. Obviously, technological differences exist between historical systems and modern-day systems, mainly due to advanced hardware capabilities. While a real-time convolution engine needed to be a specialized piece of equipment in the early 1990s [3], it can now be accomplished by freely available, open-source software tools and regular laptops [4–6], opening up many opportunities for experimentation and application. But aside from technological advances, we argue that the main difference between previous virtual acoustics research and the topics discussed in this thesis is their objective. Earlier virtual acoustic systems were often evaluated by comparing their output against a high-quality reference rendering, with the goal of maximizing similarity in direct comparison. However, even today, complete indistinguishability under all conditions is not achieved — at least not in practically relevant systems. When attempting to apply virtual acoustic technologies in practice, two questions naturally arise.

The first one is why audible differences occur. Gaining an understanding of this requires a detailed analysis of specific rendering methods. Apart from doing derivations and experiments evaluating objective metrics, listening experiments directly comparing a rendering to a high-quality reference can be used as a diagnostic tool to determine the kinds of perceptual impairments that the methods introduce. As examples, this thesis studies two properties of the spatial decomposition method (SDM). We argue that errors are mostly due to mismatches between modeling assumptions and the real world. Also, we provide arguments for why such errors are principally hard to avoid in 6 degrees-of-freedom (6DoF) rendering.

The second question is how practical virtual acoustic rendering systems for AR should be evaluated. We argue that here, the goal should not be indistinguishability in direct comparison and that evaluation paradigms must correspond closely to the goals of a system's specific application. The AR telepresence application provides us with the new, clearly defined goal of creating stable *auditory illusions* [7], i.e., virtual renderings that are believed to be real. Specifically, we introduce a paradigm for testing if virtual sound sources evoke auditory illusions even if other, real sound sources are present. We believe that this scenario is the closest to the AR telepresense applications. We call this experimental paradigm *transfer-plausibility* (TP) and define

> A system is transfer-plausible, if virtual sound sources are *believed to be real*, in the presence of real sound sources.

We argue that TP might be achievable, even if complete indistinguishability between real and virtual sources, defined as *authenticity* [8], is not reached. In turn, the goal of reaching TP leads to a specific set of requirements for the rendering system. Some potential technological building blocks of

such systems are discussed and studied here, but implementing a fully operational, 6DoF AR rendering system that reaches TP was not the goal of the thesis. Consequently, ample opportunity for future work remains, which we hope to aid and encourage with the following contributions.

Contributions Apart from presenting transfer-plausibility and discussing the evaluation of AR audio, the thesis deals with technological foundations for AR audio systems and studies some of their fundamental properties and limitations. Some are general technologies of virtual acoustics, such as parametric spatial room impulse response (SRIR) analysis and rendering. Others are more specific to sound in AR, such as blind SRIR estimation and the evaluation of transparent headphones. The topics of the publications can be summarized as follows.

– Publication I deals with a basic property of SRIR analysis: the behavior of the pseudo intensity vector in late reverberant fields.
– Publication II deals with a specific aspect of SRIR rendering: roughness perceived when rendering very transient sounds using the spatial decomposition method.
– Publication III presents a simple idea for blind SRIR estimation based on creating a pseudo-reference signal
– Publication IV presents ways of assessing headphone transparency based on auditory models. The output is correlated to the results of listening experiments.
– Publication V first introduced the idea of transfer-plausibility.
– Publication VI gives a precise definition of TP and the associated experimental design. Moreover, it compares TP to other paradigms that test for the emergence of auditory illusions.
– Publication VII allows insight into the perceptual relevance of position-dependent room acoustic differences in a self-localization task.

Background on technologies for binaural virtual acoustic rendering discussed in Publication I, and II is provided in Chapter 2. Technologies more specific to sound in AR, like blind room estimation and transparent headphones that are the topic of Publication III and IV are summarized in Chapter 3. Finally, perceptual aspects and the notion of transfer-plausibility are presented in Chapter 4 and in Publication V, VI, VII. Chapter 5 puts the contributions into context and concludes the work.

2. Virtual Acoustic Rendering

The following chapter revisits some of the existing technologies required for dynamical, binaural rendering. As such, they are important prerequisites for sound in AR telepresence but are not limited to this application. Furthermore, the chapter provides background for Publication I and II, the two studies regarding the fundamental properties and limitations of SDM.

In the realm of virtual acoustics, the process of rendering sound audible from simulated, measured or synthesized data is often referred to as *auralization* [9]. For the auralization of a specific room, spatial room impulse responses (SRIR) need to be simulated or measured, processed, and ultimately convolved with anechoic source signals. The resulting signals are then fed to loudspeaker arrays or headphones. In case head-tracking data are available, headphone rendering can be adapted according to the listener's orientation (3DoF), or position and orientation (6DoF). Such dynamical virtual acoustic rendering systems were also called *interactive virtual acoustic environments* in the past.

The first auralization techniques have already been implemented several decades ago, following first ideas by Schroeder in 1963 [10]; for a review of systems introduced before 1993, see [11]. Yet most early systems were static, i.e., they did not include head-tracking. One of the first systems to enable dynamical binaural auralization with head-tracking was the *convolvotron*, which allowed time-varying convolution by means of a dedicated processor [3]. Some examples of room acoustic rendering and auralization software developed over the past decades were the Binaural Room Simulation software in [12], DIVA [13, 14], SLAB [15], RAVEN [16], or RAZR [17].

2.1 Simulation, Measurement, Estimation

As the goal of auralization systems often was to aid room acoustic design [18], many early systems take room geometry, wall materials, and source/receiver positions as input and render a simulation of the room.

They were frequently based on geometrical acoustics [19] and to a limited extent on wave-based simulation and hybrids of the two (see [20, 21] for recent introductions to wave-based room acoustic simulation). When the acoustics of an existing room shall be reproduced, techniques based on measured SRIR are often more appropriate. This is because simulations would require an exact model of the room including the acoustical properties of all materials. The accuracy of the simulation might suffer from modeling errors or insufficiencies of the simulator itself. Measurements of existing spaces are used as references for testing simulations, as in the round robin described in [22]. Also, the work presented in this thesis mostly uses measurements. To obtain a SRIR, a measurement signal is played back from a loudspeaker and recorded using one or more microphones in a room. Exponential sine sweep [23, 24] are commonly used as measurement signals. After recording, the microphone signals are deconvolved with the measurement signal, and one RIR is obtained for every microphone channel.

We define a SRIR as consisting of several RIRs from different microphones at different positions, which provides spatial information. A SRIR could be measured using binaural microphones, where we speak of a binaural room impulse response (BRIR) or different microphone arrays. If the microphone array is compact, the term directional room impulse response (DRIR) is used, too. Most binaural auralization methods can be applied to both measured or simulated SRIRs. Such methods have been used in listening experiments that are part of this thesis, e.g., in Publication II, IV, V, VI, VII. For practical AR systems, simulations or measurements are likely to be replaced by estimates, which is discussed further in Section 3.1.

2.2 Creating Binaural Auralizations

Some of the most fundamental approaches for binaural auralization of rooms are summarized in the following section. More detailed treatments of the topics can be found in [25–28], for example.

2.2.1 Binaural Room Impulse Responses

Conceptually, the simplest system for creating a binaural rendering of an existing room is to measure a BRIR using microphones in the ears of a human listener or an artificial head and torso simulator (HATS). A BRIR consists of two impulse responses, $b_{\mathrm{L}}(t)$, and $b_{\mathrm{R}}(t)$, one for the left and one for the right ear, where t denotes time. An example is shown in Fig. 2.1. Once a BRIR is obtained, a virtual sound source can be reproduced using headphones, by convolving the BRIR with an anechoic, monophonic input

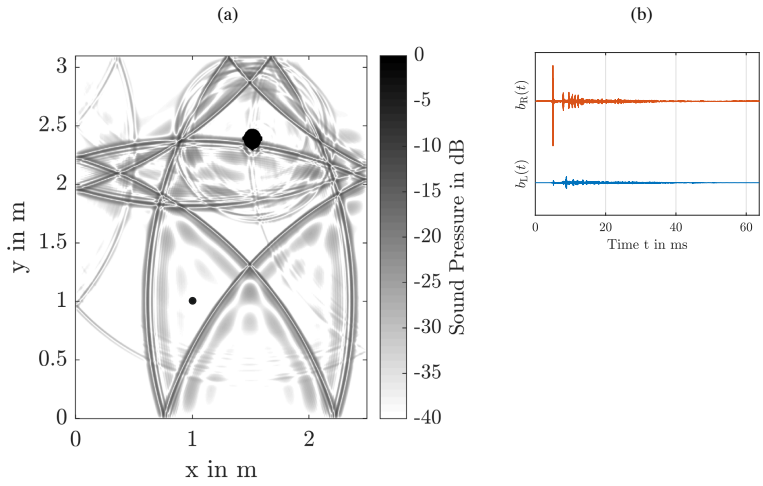


Figure 2.1. (a) Visualization of the sound pressure in a room with a head in it, $t = 10$ ms after impulse excitation at the position of the small black circle in the lower left. After the excitation, sound is reflected from the walls and the head. (b) Left and right channels of the BRIR measured at the ear positions, where the direct sound is seen with a higher level and earlier arrival in the right channel response and several reflections are visible in the response of both ears.

signal $s(t)$ to produce the binaural rendering $\boldsymbol{y}(t)$, as in

$$\boldsymbol{y}(t) = \begin{bmatrix} b_\mathrm{L}(t) \\ b_\mathrm{R}(t) \end{bmatrix} * s(t). \tag{2.1}$$

In addition, headphone compensation should be employed, for which different approaches exist [29–31]. Ideally, headphones are measured using the same binaural receiver. Such a minimal, static auralization method was used for certain tests in Publication II, IV, and VII, where specific acoustic aspects of a room at one position and orientation were of interest.

If head-tracking shall be taken into account, the BRIR needs to be measured for all possible head orientations, so that the response with the orientation closest to the head orientation of the user can be loaded into memory upon rendering. An open-source implementation of such dynamic BRIR-based rendering is offered, for example, by the *pyBinSim* toolbox [4]. It uses partitioned convolution, studied in detail by [32]. The disadvantage of BRIR-based 3DoF rendering techniques is that they requires a large number of measurements to be taken and stored and that the specific binaural cues of the employed binaural receiver, be it a human subject or a HATS, are inseparable from the room response. More flexibility is offered by techniques employing microphone arrays.

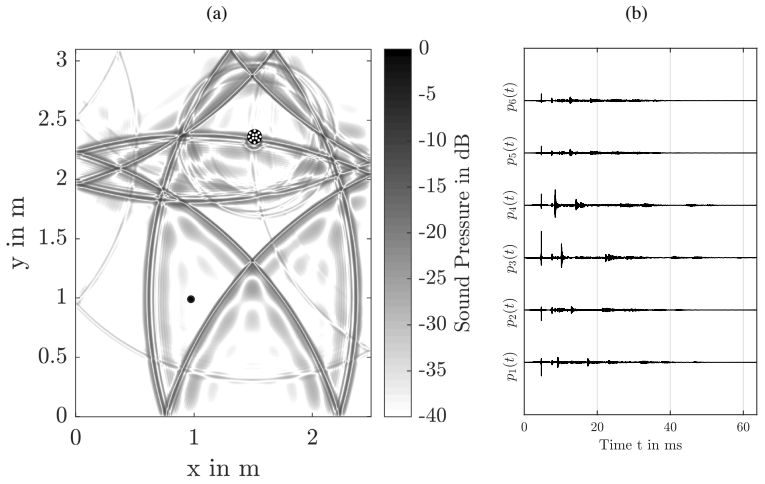


Figure 2.2. A simulation of a microphone array in a room, using the same method and parameters as in Fig. 2.1. Now (a) shows the room with a schematic, rigid sphere microphone array, (b) shows the response of six microphone channels. Direct sound and reflections arrive at the microphone capsules at different levels and at different times.

2.2.2 Microphone Array Measurements

Instead of using binaural receivers, microphone arrays can be used to obtain microphone array RIRs, illustrated in Fig. 2.2. Different array designs are available; Fig. 2.3 shows some examples like an open array with omnidirectional capsules and one with cardioid capsules, as well as a rigid array with omnidirectional capsules. All of these designs are considered spherical microphone arrays (SMAs) [33], as the microphone capsules lie on a sphere. In AR practice, non-spherical designs are also relevant, as arrays of microphones placed on glasses or headsets will often need to be used, as employed in [34–36], for example.

An M-channel microphone array RIR can be denoted as

$$\boldsymbol{p}(t) = \begin{bmatrix} p_1(t) \\ p_2(t) \\ \vdots \\ p_M(t) \end{bmatrix}, \tag{2.2}$$

where $p_m(t)$ are the responses measured using M microphone capsules. A microphone array response can not be auralized directly. It needs to be processed to obtain a loudspeaker array or a binaural signal.

(a) G.R.A.S. 3D Probe **(b)** SoundField ST450 **(c)** Eigenmike EM32

Figure 2.3. Different commercially available spherical microphone arrays, utilized in different publications in this thesis.

2.2.3 Head-Related Transfer Functions

When the objective is to render binaural signals based on microphone array measurements, head-related transfer functions (HRTFs) are required. They are measured just as BRIRs, by recording a measurement signal with binaural microphones, but are performed in an anechoic room. HRTFs are measured using a grid of source directions, obtainable through setting up an arc of loudspeakers and rotating the subject on a turning chair, or through rotating the loudspeaker arc around the subject. For fast measurements of many directions, interleaved exponential sine sweeps can be used [37]. Plenty of comprehensive literature on HRTF acquisition, analysis, and processing is available, e.g., [38, 39]. Also, several datasets of HRTF measurements are openly available, such as the ARI database [40], HUTUBS [41], SONICOM [42], 3D3A [43], SADIE II [44], or the ITA HRTF [45]. Nowadays, HRTFs are often stored in the *spatially oriented format for acoustics* (SOFA), which is standardized by the Audio Engineering Society. Currently, the most recent version is AES69-2022, details of which are described in [46].

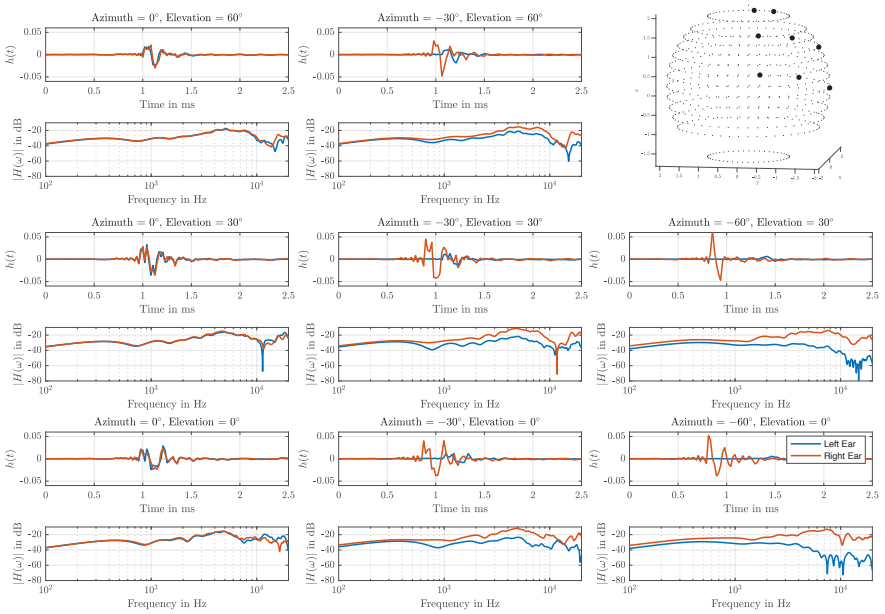


Figure 2.4. HRIRs and HRTFs of the author, measured using PIRATE [1] microphones in the multichannel playback room "Wilska" at the Aalto Acoustics Laboratory. The selected directions on the sphere are indicated in the upper right by the big black dots. All measured directions are indicated by the small dots.

A set of HRTFs is a collection of $2 \times L$ transfer functions measured at directions $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \dots \boldsymbol{\theta}_L\}$ around a subject

$$\boldsymbol{H}(\boldsymbol{\Theta}, \omega) = \begin{bmatrix} h_\mathrm{L}(\boldsymbol{\theta}_1, \omega) & \dots & h_\mathrm{L}(\boldsymbol{\theta}_L, \omega) \\ h_\mathrm{R}(\boldsymbol{\theta}_1, \omega) & \dots & h_\mathrm{R}(\boldsymbol{\theta}_L, \omega) \end{bmatrix}, \tag{2.3}$$

where ω denotes frequency. Fig. 2.4 shows some examples of HRTFs and their time domain representations, called head-related impulse responses (HRIRs).[1] These examples show that the further the sound source moves to the right (negative azimuth angles), the higher the level in the right ear in relation to the level at the left ear (Interaural Level Differences, ILDs). Furthermore, the sound arrives earlier at the right ear (Interaural Time Differences, ITDs). Between the different elevations, spectral differences are seen at high frequencies. These differences are highly individual and if there is a mismatch between a person's own HRTF and an HRTF used for rendering, vertical localization performance is decreased and front-back confusions become more likely [47]. Therefore, the topic of HRTF individualization is often discussed. However, applications may exist in which individualization is not required. Also, research has shown that adaptation to a non-individual set of HRTFs is possible to some

[1] Note that in the following, we distinguish between time and frequency domain quantities only by changing the argument, e.g., $h(\omega)$ is the Fourier transform of $h(t)$.

extent, as reviewed in [48]. A recent initiative [49] plans on fostering long-term adaptation by standardizing one particular set of HRTFs that all manufacturers would use as a default whenever no individualization is employed. Like this, users would get the chance to adapt to one set over a long period of time and across different applications.

A set of HRTFs alone can be used for rendering sound sources without room information. Therefore, interpolation between HRTFs is required. One common choice is to create a triangular mesh from the measurement grid and to weight and add the HRTFs of the triangle that contains the desired sound source direction. With a particular choice of weights, this procedure is equivalent to vector base amplitude panning (VBAP) [50]. As interpolating HRTFs directly can cause coloration through comb-filtering, improved methods exist. A common option is to first estimate the time delays contained in the HRTF using one of the methods compared in [51], and to then design a minimum-phase filter with the same magnitude response as the HRTF. The minimum-phase filters are then interpolated, as well as the time delay values. Finally, the required time delay is re-introduced. This approach has been used for a long time [13, 52, 53], and is also implemented in [5], which was used in Publication V. Recently, it has been shown to be improvable by additional magnitude correction [54].

2.2.4 Direct Binaural Decoding using HRTFs

With a set of HRTFs available, microphone array room impulse response measurements can be converted to BRIRs. A simple method is to determine a set of filters that minimize the squared difference between the array directivity pattern and the HRTF at each frequency [55–58]. In addition to the HRTFs and the microphone array RIR to be rendered, anechoic microphone array measurements are required. For simplicity, we assume that microphone array measurements were conducted at the same positions as the HRTF measurements, $\boldsymbol{\Theta}$, so that

$$\boldsymbol{A}(\boldsymbol{\Theta},\omega) = \begin{bmatrix} a_1(\boldsymbol{\theta}_1,\omega) & \dots & a_1(\boldsymbol{\theta}_L,\omega) \\ \vdots & & \\ a_M(\boldsymbol{\theta}_1,\omega) & \dots & a_M(\boldsymbol{\theta}_L,\omega) \end{bmatrix}. \tag{2.4}$$

is a set of so-called array transfer functions (ATFs). Now, a filter is sought that, when applied to a microphone response, converts it to a binaural response

$$\boldsymbol{b}(\omega) = \boldsymbol{C}(\omega)\boldsymbol{p}(\omega) \tag{2.5}$$

$$= \begin{bmatrix} \boldsymbol{c}_{\mathrm{L}}^{\mathrm{T}}(\omega) \\ \boldsymbol{c}_{\mathrm{R}}^{\mathrm{T}}(\omega) \end{bmatrix} \boldsymbol{p}(\omega), \tag{2.6}$$

where $\boldsymbol{c}_\mathrm{L}(\omega)$ and $\boldsymbol{c}_\mathrm{R}(\omega)$ are the transfer functions of filters converting the microphone responses to left and right binaural responses, and $.^\mathrm{T}$ denotes transposition. Note that the following operations are carried out in the frequency domain, for each frequency ω separately, but that the dependency on frequency is dropped in the following equations for better readability.

The filter shall combine the microphone response so that the resulting directivity patterns match those of the HRTFs in the least squares sense

$$\boldsymbol{C}^\mathrm{LS} = \arg\min_{\boldsymbol{C}} \|\boldsymbol{C}\boldsymbol{A} - \boldsymbol{H}\|_\mathrm{F}^2, \tag{2.7}$$

where $\|.\|_\mathrm{F}$ denotes the Frobenius norm. This can be achieved using the Moore-Penrose pseudoinverse, defined as $\boldsymbol{A}^\dagger = \boldsymbol{A}^\mathrm{H}(\boldsymbol{A}\boldsymbol{A}^\mathrm{H})^{-1}$

$$\boldsymbol{C}^\mathrm{LS} = \boldsymbol{H}\boldsymbol{A}^\dagger, \tag{2.8}$$

where $.^\mathrm{H}$ is the hermitian transpose and $.^{-1}$ denotes matrix inversion. In practice, a regularisation constant β can be used to limit the allowed amplifications, which would otherwise lead to excessive boosts of sensor noise

$$\boldsymbol{C}^\mathrm{reg} = \boldsymbol{H}\boldsymbol{A}^\mathrm{H}(\boldsymbol{A}\boldsymbol{A}^\mathrm{H} + \beta\boldsymbol{I})^{-1}. \tag{2.9}$$

Different regularization approaches were compared in [57]. A limitation of this approach is that a unique pair of filters needs to be computed (and stored) for all possible head orientations of the listener [58].

2.2.5 Ambisonics Processing

Another method for auralizing microphone array signals relies on the principles of Ambisonics [59–61]. Ambisonics processing is more versatile in that HRTF and ATF measurement grids can be different, and a rendered signal can efficiently be rotated to accommodate the listeners' orientation. The microphone array responses are first *encoded* into the spherical harmonics (SH) domain. There, rotations of the entire field are easy to realize and the rotated field can then be *decoded* to a binaural signal.

After the encoding stage, each channel of the response represents one SH pattern. In acoustics, real SHs, as represented in Fig. 2.5, are usually defined as

$$Y_n^m(\boldsymbol{\theta}) = N_n^{|m|} P_n^{|m|}(\cos\vartheta) \begin{cases} \sqrt{2}\sin(|m|\phi) & m < 0 \\ 1 & m = 0 \\ \sqrt{2}\cos(m\phi) & m > 0 \end{cases}, \tag{2.10}$$

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

where $Y_n^m(\boldsymbol{\theta})$ is the spherical harmonic of order $n \le N$ and degree $-n \le m \le n$ evaluated at $\boldsymbol{\theta} = \begin{bmatrix} x & y & z \end{bmatrix}^\mathrm{T} \in \mathcal{S}^2$ which is a direction vector on the

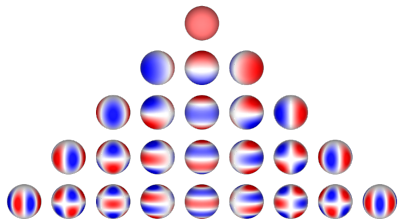


Figure 2.5. Real spherical harmonics up to order $N = 4$.

unit sphere that can also be represented using the azimuth angle ϕ and elevation angle ϑ. P_n^m are the associated Legendre polynomials.

For convenience and practical implementations, it is useful to define a matrix/vector notation for SH processing. The most common way to stack the spherical harmonics Y_n^m into a vector uses Ambisonics Channel Numbering (ACN), which is defined in the AmbiX format [62]. The index for each SH component of order n and degree m is found by $i = n^2 + n + m$. Like this, a vector of spherical harmonics can be written as

$$\boldsymbol{y}_N(\boldsymbol{\theta}) = \begin{bmatrix} Y_0^0(\boldsymbol{\theta}) & Y_1^{-1}(\boldsymbol{\theta}) & Y_1^0(\boldsymbol{\theta}) & Y_1^1(\boldsymbol{\theta}) & \dots & Y_N^N(\boldsymbol{\theta}) \end{bmatrix}^{\mathrm{T}}, \tag{2.11}$$

SHs evaluated at several directions can be stacked into a matrix as in

$$\boldsymbol{Y}_N(\boldsymbol{\Theta}) = \begin{bmatrix} \boldsymbol{y}_N(\boldsymbol{\theta}_1)^{\mathrm{T}} \\ \boldsymbol{y}_N(\boldsymbol{\theta}_2)^{\mathrm{T}} \\ \vdots \\ \boldsymbol{y}_N(\boldsymbol{\theta}_L)^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} Y_0^0(\boldsymbol{\theta}_1) & Y_1^{-1}(\boldsymbol{\theta}_1) & Y_1^0(\boldsymbol{\theta}_1) & \dots & Y_N^N(\boldsymbol{\theta}_1) \\ Y_0^0(\boldsymbol{\theta}_2) & Y_1^{-1}(\boldsymbol{\theta}_2) & Y_1^0(\boldsymbol{\theta}_2) & \dots & Y_N^N(\boldsymbol{\theta}_2) \\ \dots & & & & \\ Y_0^0(\boldsymbol{\theta}_L) & Y_1^{-1}(\boldsymbol{\theta}_L) & Y_1^0(\boldsymbol{\theta}_L) & \dots & Y_N^N(\boldsymbol{\theta}_L) \end{bmatrix}. \tag{2.12}$$

Encoding A filterbank of encoding filters can be found from the ATFs. The process is analogous to the direct decoding approach. Only now, the target is not the binaural directivity patterns, but the SHs [63]

$$\boldsymbol{E} = \boldsymbol{Y}_N^{\mathrm{T}} \boldsymbol{A}^{\mathrm{H}} (\boldsymbol{A}\boldsymbol{A}^{\mathrm{H}} + \beta \boldsymbol{I})^{-1}, \tag{2.13}$$

where $\boldsymbol{Y}_N$ is a matrix of the SHs up to order N, evaluated at the ATF measurement positions $\boldsymbol{\Theta}$, and $\boldsymbol{E}(\omega)$ is a multiple-input, multiple-output (MIMO) filterbank with $(N+1)^2 \times M$ different filters. The encoder is then applied to yield the SH domain room response. In the frequency domain, this can be written as

$$\breve{\boldsymbol{h}}(\omega) = \boldsymbol{E}(\omega)\boldsymbol{p}(\omega), \tag{2.14}$$

where the inverse Fourier-Transform of $\breve{\boldsymbol{h}}(\omega)$ is the n-th order SH domain RIR, referred to as SH-RIR or Ambisonics RIR (ARIR).

In the special case of a spherical microphone array, this MIMO system can be factored into an $(N+1)^2 \times M$ matrix of real coefficients $\bar{\mathbf{E}}$ and a bank

of only N unique filters [33, 61]. In the frequency domain, the encoding process with such a system can be denoted as

$$\check{\boldsymbol{h}}(\omega) = \mathrm{diag}_N\{f_n(\omega)\}\bar{\mathbf{E}}\boldsymbol{p}(\omega), \tag{2.15}$$

where diag_N creates a diagonal matrix that repeats the n-th order coefficients $2n+1$ times.

Once an SH-SRIR is obtained, it can be convolved with a monophonic, anechoic input signal to create the Ambisonics signal $\check{\boldsymbol{x}}$

$$\check{\boldsymbol{x}}(t) = \check{\boldsymbol{h}}(t) * s(t). \tag{2.16}$$

To account for listener orientation, this entire "scene" can be rotated by means of matrix multiplication with the appropriate rotation matrix [64]

$$\check{\boldsymbol{x}}_{\mathrm{rot}}(t) = \boldsymbol{R}_N(-\gamma, -\beta, -\alpha)\check{\boldsymbol{x}}(t), \tag{2.17}$$

where α, β, γ are Euler angles describing the head orientation of the listener.

Binaural Decoding After possible rotation, the signal is decoded using a so-called binaural decoder. A simple decoder $\boldsymbol{D}^{\mathrm{LS}}$ is the least squares decoder, for which the difference between the SH patterns and the HRTF is minimized in the least squares sense

$$\boldsymbol{D}^{\mathrm{LS}} = \arg\min_{\boldsymbol{D}} \left\| \boldsymbol{Y}_N\boldsymbol{D} - \boldsymbol{H}^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \tag{2.18}$$

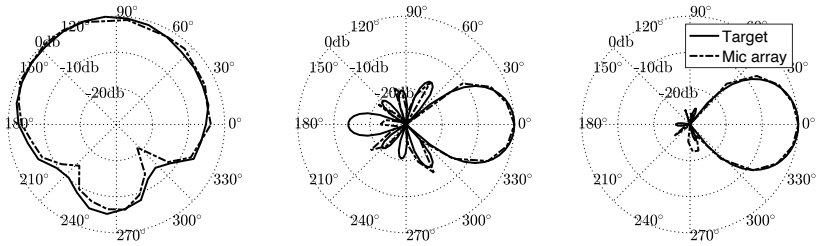
$$\boldsymbol{D}^{\mathrm{LS}} = \boldsymbol{Y}_N^{\dagger}\boldsymbol{H}^{\mathrm{T}}. \tag{2.19}$$

When using low orders, e.g., $N = 1$, such a decoder yields inaccurate localization and loss of high frequencies. Partially, this can be compensated by diffuse field equalization [65]. What has been shown to be highly beneficial is to modify the phase of the HRTFs when creating the decoder [66]. While time differences at high frequencies increase the order that is required to represent the set of HRTFs, they do not contribute strongly to localization. Hence, removing time differences above a certain cut-on frequency (e.g., 1.5 kHz) through time- or phase alignment [66] reduces the required order for achieving a good magnitude fit while not reducing the perceptual localization accuracy. Another approach is to formulate the problem as a *magnitude least squares* problem above the cut-on, effectively simplifying the phase as well [67–69].

Beamforming No matter how the weights are found, each of the two rows of $\boldsymbol{D}$ can also be interpreted as one *SH-domain beamformer* (sometimes called "modal" beamformer) in that it takes the SH signals and weights and combines them to create a certain directivity pattern; in the case of binaural decoding, this pattern approximates the directivity of a human ear. In general, a beamformer can be written as

$$x_{\mathrm{s}}(\omega) = \boldsymbol{d}^{\mathrm{T}}(\omega)\check{\boldsymbol{x}}(\omega), \tag{2.20}$$

where $\boldsymbol{d}(\omega)$ is a vector of coefficients. Stacked into matrix $\boldsymbol{D}(\omega)$, multiple beamformer outputs can be obtained, as in the case of the binaural decoder. A special case is axis-symmetric SH-domain beamforming, where the coefficients can be real, frequency-independent, and equal for all $2n+1$ components belonging to each order n., i.e., one row in Fig. 2.5. Specific axis-symmetric beamformer designs are often found by globally optimizing a certain objective, such as the directivity index (DI) or the length of the $\boldsymbol{r}_\mathrm{E}$ vector. Fig. 2.6 shows the directivity of a left ear at 1 kHz and these two beamforming patterns with $N = 4$ as solid lines. The dotted lines show the patterns achieved based on encoding microphone array signals from an Eigenmike em32. A max-$\boldsymbol{r}_\mathrm{E}$ beamformer based on em32 signals is used in Publication III.



(a) Left ear at 1 kHz. **(b)** max-DI **(c)** max-$\boldsymbol{r}_\mathrm{E}$

Figure 2.6. Target beampatterns as solid lines and achieved beampatterns using an Eigenmike em32 as dashed lines. Encoding filters and plots are based on measurements from [2].

Real-time implementations of Ambisonics-based processing are available through combinations of VST plugins for example from [5] or [70], or through the python toolbox introduced in [71, 72]. Real-time binaural decoding of Ambisonics signals is an appropriate approach for 360° videos in VR, like the production we presented in [73]. So far, *YouTube* implemented first-order Ambisonics, whereas *HOAST*[2] allows for up to fourth order [74].

2.2.6 Parametric SRIR Methods

The main idea of parametric SRIR processing is to employ a general SRIR model, which is parameterized using a specific measurement. Tab. 2.1 gives an overview of existing methods. All methods consist of a measurement stage, an analysis stage, in which parameters are computed, and a synthesis stage, in which responses for arbitrary reproduction setups like loudspeaker arrays or head-tracked headphones are created. In addition, once spatial parameters are computed, they can be used for technical

[2] `https://hoast.iem.at/`

Method	SIRR [81]	HO-SIRR [78]	SDM [77]	ASDM [82]	*Stade* [83]	RSAO [84]	REPAIR [85]
Measurement	FO SMA	HO SMA	Open Array	FO SMA	Very HO SMA	Circular	Any
Analysis	Narrowband PIV, Diffuseness	Narrowband Sector PIV, Sector Diffuseness	Broad- [77] or Narrow-band [86] TDoA	Broadband PIV	SRP, Peak Detection	SRP, Peak Detection	MUSIC
Synthesis	VBAP, Diffuse Rendering	VBAP, Directional Diffuse Rendering	NLS [87] or VBAP [77], Spectral Correction [88]	Ambisonics, Spectral Correction	Binaural, Parametric Diffuse Rendering, coherence matching	Object-based, Parametric Diffuse Rendering	Any

Table 2.1. Structure of SRIR processing algorithms, consisting of measurement, analysis, and synthesis, and methods with their specific processing choices.

analyses of the response, as done in [75] and [76] for concert halls, and in Publication VI and VII for the variable acoustics room used therein.

A multitude of parametric SRIR methods has been introduced over the years, and to this point, it is unclear which provides the best results in which situation. Although direct comparisons between methods have been conducted in previous research [77–80], with many inter-dependencies on specific microphone array choices, test signals, and testing methods, it is not surprising that results disagree considerably between experiments. All parametric methods have in common the reliance on a SRIR model, for which they estimate parameters.

Spatial Decomposition Method Undoubtedly, the simplest sound field model underlies SDM [77]. It assumes that the sound field consists of a set of plane-waves arriving at the receiver array. Exactly one plane-wave is assumed to arrive at each time instance. Assuming an ideal, first-order SH-RIR, the model can be written as

$$\check{\boldsymbol{h}}(t) = p(t)\boldsymbol{y}_1(\boldsymbol{\theta}(t)). \tag{2.21}$$

As the pressure $p(t)$ is readily available from omnidirectional capsules or the 0th order SH channel, the only parameter to estimate is the direction of arrival at each time instance, $\boldsymbol{\theta}(t)$. This can be done through several directional estimators. Two possible methods were already proposed in early work regarding SRIR analysis [89]: Time Difference of Arrival (TDoA) estimation and pseudo intensity vector (PIV) estimation.

The method first presented in the context of SDM was TDoA estimation [77] (see full equations in [75, 90]). A method studied in more detail by the author is the instantaneous PIV. It was also used in [91–93], where the SDM variant employing it is called Ambisonics SDM (ASDM).

The PIV is straightforward to compute from an SH-RIR, as

$$\mathbf{PIV}(t) = \check{h}_\mathrm{w}(t) \begin{bmatrix} \check{h}_\mathrm{x}(t) \\ \check{h}_\mathrm{y}(t) \\ \check{h}_\mathrm{z}(t) \end{bmatrix} = \check{h}_0(t) \begin{bmatrix} \check{h}_3(t) \\ \check{h}_1(t) \\ \check{h}_2(t) \end{bmatrix}, \tag{2.22}$$

where the indexing in the second expression corresponds to the ACN convention as in Eq. (2.11).

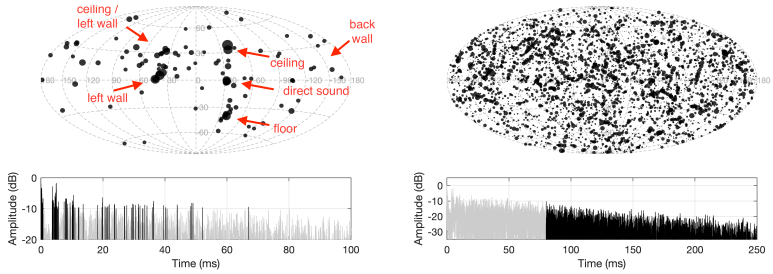
In the early part of the response, the PIV points to the direction of individual early reflections. This can be seen on the left of Fig. 2.7 for three different room configurations of the variable acoustics room “Arni”. The room has been used extensively in several studies, e.g., Publication I, VI, and VII, and is described in more detail in [94]. Both the direct sound as well as the left wall reflection, and several other reflections are visible in the analysis.

However, the reflection density quickly increases. In a shoebox-shaped room, for example, the number of reflections per time instance has been shown to increase quadratically [95]. Consequently, it becomes impossible to estimate individual reflections later in the response correctly. The reason for this failure is a discrepancy between modeling assumptions and reality: more than one reflection occurs per time instance.

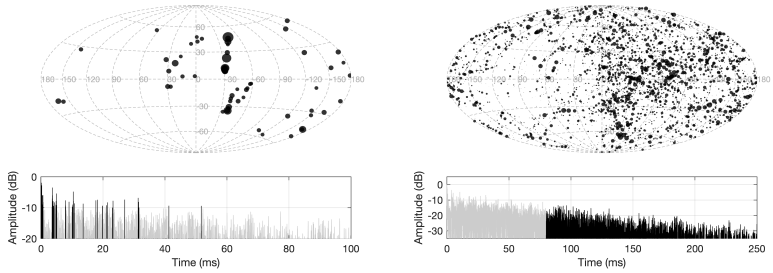
Nevertheless, it is easy to see in Fig. 2.7 that less energy is assigned to the left when the left wall was made absorptive and also the back when that wall was made absorptive as well. To explain this behavior, Publication I derives the directional distribution of the PIV in anisotropic late fields, as they are studied in [96–98], for example. For the derivation, a variant of the PIV is used, for which the first-order responses are divided rather than multiplied by the pressure, as in

$$\hat{\boldsymbol{\theta}}(t) = \frac{1}{\check{h}_\mathrm{w}(t)} \begin{bmatrix} \check{h}_\mathrm{x}(t) \\ \check{h}_\mathrm{y}(t) \\ \check{h}_\mathrm{z}(t) \end{bmatrix} = \frac{1}{\check{h}_0(t)} \begin{bmatrix} \check{h}_3(t) \\ \check{h}_1(t) \\ \check{h}_2(t) \end{bmatrix}. \tag{2.23}$$

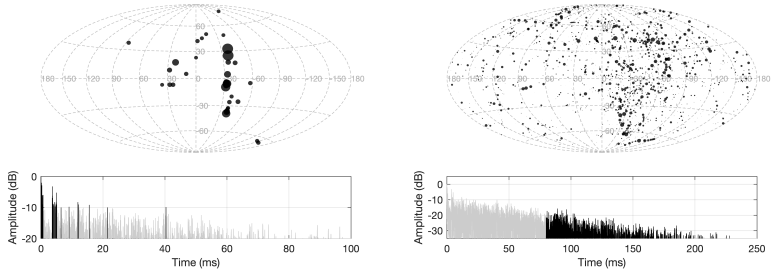
This variant could be called “self-normalized” or “pressure-normalized” PIV, or it could be seen as a special case of the generalized time-domain velocity vector [99]. In the presence of a single plane wave, it has unit length. Independent of the field, it points to the same direction as the PIV defined in Eq. (2.22). For the derivation in Publication I, the pressure-normalized PIV was used, because it most easily led to an analytical solution of the directional distribution in a late field. The main result is that the directional distribution of $\hat{\boldsymbol{\theta}}$ follows the directional distribution of the late field to some extent. The result can predict that it is less likely to obtain directional estimates from the directions of the absorbing walls in the late field as it is seen in Fig. 2.7. However, the results also clarify that it can not capture complicated spatial variation, as the resulting



(a) All walls in a reflective setting.



(b) Left wall in absorbing setting. The first-order left wall reflection is weaker, also in the late part there is less energy from the left.



(c) Left and rear wall in absorbing setting. The first-order left wall reflection is weaker, the back-wall reflection is not seen at all. Also, in the late part, there is less energy from the left and from the back.

Figure 2.7. Analysis of early (0 – 80 ms) and late parts (80 ms – 250 ms) of the response in Arni in three different room acoustical settings.

expression has limited degrees of freedom. Thus, if the PIV is used as the basis for parametric rendering in AR audio, one needs to consider that the directionality of the late field is not ideally captured in all situations. In [100], we show a listening experiment, which demonstrates that under ideal conditions, such late-field directionality is, in principle, audible. As a consequence, neglecting it may make a rendered response distinguishable from a reference in some situations.

Note that while Publication I only studies the properties of PIV estimation, [90, 100] also show observed directional distributions for TDoA estimation.

SDM Rendering Once the DoAs are estimated, SDM rendering uses them to synthesize loudspeaker array responses, or binaural responses. The simplest algorithm employs nearest loudspeaker synthesis (NLS) [77], where each sample of the omnidirectional response is simply mapped to one of the loudspeaker responses g_l used for rendering, by the windowing functions $w_l(t)$

$$g_l(t) = w_l(t)p(t), \tag{2.24}$$

with

$$w_l(t) = \begin{cases} 1 & l = l_{NL}(t) \\ 0 & l \neq l_{NL}(t), \end{cases} \tag{2.25}$$

$$l_{\mathrm{NL}}(t) = \arg\min_l \|\hat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}_l\|, \tag{2.26}$$

where $\boldsymbol{\theta}_l$ are the directions of the L loudspeakers. As a consequence of this assignment, all reproduction channel responses still sum up to the original pressure response, if added coherently, i.e.,

$$\sum_l w_l(t) = 1. \tag{2.27}$$

The summation of the reproduction channel responses at the listener's ears, however, is not equal to the original pressure. If the loudspeaker responses are convolved with HRIRs to create a binaural rendering or a listener is assumed to be sitting in a loudspeaker array, the ear signals are instead equal to

$$\boldsymbol{b}(t) = \sum_{l=1}^{L} g_l(n) * \boldsymbol{h}_l(t), \tag{2.28}$$

where $\boldsymbol{h}_l$ is the HRTF from the direction of the loudspeaker l. Now, the head as a scattering object of finite size prevents perfect summation, as it is further analyzed in Publication II. The main consequence is that some properties of the sparse reproduction loudspeaker responses are maintained in the binaural signals, which in turn leads to two perceptual effects.

First, if very transient sounds are rendered, an effect described as "roughness" or "graininess" can be perceived. The effect has been studied in detail in Publication II. Since then, methods to compensate for it have been proposed: In [101], binaural responses are convolved with a short allpass sequence to increase temporal density. In [102], additional pulses are assigned to other directions than the estimated one, based on the output of four beamformers directed towards the vertices of a tetrahedron, rotated such that one of them points to the estimated DoA.

Second, the rendered responses have excess high-frequency content. This phenomenon has been noticed early on in the development of SDM, and corrections based on spectrogram inversion [88], octave-band envelope correction [93], or exponential slope correction [101] have been proposed. In listening experiments, SDM rendering has been shown to reach relatively high levels of plausibility [101, 103] and to reproduce important perceptual features of the response. However, renderings are usually distinguishable from a BRIR reference rendering [80, 100, 104].

Note that correcting these two effects post-rendering is the only possible option, as the effect is inherent to the modeling assumptions of SDM. Any attempt to reduce them by constraining spatial assignment, using fewer loudspeakers or the like, would interfere with the functioning principle itself.

Multidirectional Analysis and Rendering Other methods (summarized in Table 2.1) assume more complex sound field models than SDM. A much more advanced sound field model is adopted by higher-order spatial impulse response rendering (HO-SIRR) [78], which is an improved version of SIRR [81]. The main difference to SDM is that DoA estimation is performed in time-frequency tiles, and in multiple, spatially constrained regions. Additionally, a so-called diffuseness parameter is measured in each time-frequency tile and sector. The sector analysis also permits a more highly resolved representation of the spatial distribution in the late reverberation. Instead of using constrained sectors, the method introduced in [105] allows for multidirectional estimation with the MUltiple SIgnal Classification (MUSIC). Its sound field model is much more advanced than that of SDM, but in [105], we also demonstrate that parameterizing it is more challenging. Although HO-SIRR and the method from [105] can produce high-quality output, not all renderings are indistinguishable from a binaural reference in all situations either [105].

2.3 Measurement-Based 6DoF Systems

All the methods for analyzing and rendering measured SRIR presented so far only considered one response, measured at one configuration of source and receiver. This allows for rendering at one listener position with variable orientation. Now we will discuss the implications of allowing the listener to move as well.

As mentioned in Sec. 2.2.1, in the case of 3DoF BRIR rendering, the correct response is selected depending on the orientation of the listener. This requires a large dataset of measurements. In [106], the noticeable limit in the directional resolution of a cross-fading-based HRTF rendering system was approximately 5° for the most sensitive 5% of listeners. Assuming an equiangular sampling grid with a resolution of said 5°, 2486 binaural

measurements are needed. If, in addition, horizontal movements should be allowed in a 2×2 m area with a grid resolution of 20 cm (as in [107]), these 2486 measurements would need to be taken at 121 measurement positions so that the total number of measurements would amount to 300 806. Note that in this calculation, we have neither considered different head-above-torso positions [108], nor allowed for vertical movements, nor for multiple sources, which all would increase the number of measurements even further. Such measurement effort is rarely feasible, even in a research context.

Using the same assumption of a 2D sampling grid, but employing a rendering system based on microphone array SRIR (see Sec. 2.2.2), only 121 measurements with M channels each need to be taken, which is easier to realize already. These measurements can then be encoded to the SH domain. Upon rendering, the closest SH domain responses can then be selected depending on the listener's position and rotated according to the listener's head orientation before decoding it to a binaural signal. To realize such systems, we have developed a tool called the 6DoF convolver [6]; similar software is presented in [109]. While measuring SRIR at 121 positions is feasible in a research context, especially if robotic measurement systems are employed [110–112], using fourth-order SH domain responses would still lead to $121 \times 25 = 3025$ response to store and load. For practical AR systems with resource constraints, this is likely too much.

The conclusion of these considerations is that 6DoF rendering based on directly auralizing measurements is infeasible for most practical systems. Such systems must, therefore, rely on sound field modeling assumptions, such as those of existing parametric SRIR models introduced in Sec. 2.2.6, or other sets of physical or perceptual assumptions. A good example is [113], where certain assumptions about directional resolution and distance cues are made to minimize rendering effort. In a wider sense, even treating direct sound and reverberant tail as separate implies modeling assumptions. Furthermore, approaches introduced as interpolators crucially depend on models. In [114], for example, we tried to find a parameterization of the sound field in an extended region by employing SDM analysis at several positions and combining the data to find spatially localized "image sources" more robustly. In [115], we presented interpolation ideas for such image sources. In both papers, we focussed solely on objective measures. [116] proposed an algorithm for the analysis and position dynamic rendering of SRIRs measured at several locations. An interpolation method that was only based on mitigating perceptual consequences was proposed in [117]. Therein, microphone array measurements on a line were used. Perceived similarity to a reference rendering with 5 cm resolution was small for distances between measurements of up to 50 cm. However, no comparison to a real reference was conducted; the rendering would likely not have been indistinguishable.

What has already been mentioned in the sections above is that although many parametric models of SRIR have been proposed, along with algorithms that estimate their parameters, their rendering outputs were usually not completely indistinguishable from a reference. In Chapter 4, indistinguishability from a real-world reference is defined as authenticity [8]. Taking together the fact that 6DoF rendering necessitates models but that model-based rendering usually does not lead to authentic reproduction implies that other goals are required, like those discussed in Chapter 4.

3. Developing Virtual Acoustic Systems for AR Telepresence

The previous section discussed some of the fundamental methods and limitations of binaural auralization methods for measured or simulated SRIRs. In real-world AR telepresence applications, additional technological challenges arise. First, SRIRs usually cannot be measured but must be estimated blindly from limited information. Second, room acoustic rendering will likely not be able to rely on convolutions but must employ more efficient algorithms. And third, headphones that are transparent to real-world sound need to be employed. These technologies are discussed in the following sections. A sketch of a possible full AR telepresence rendering system is shown in Fig. 3.1. A vision of an AR system, not only allowing for telepresence, but also for modifying a users acoustic surrounding at will is presented in [118].

Historically, the idea of augmenting the auditory space with virtual sound sources has been pursued even before real-time visual rendering with glasses was within reach. It was sometimes called audio augmented reality (AAR) [119, 120]. Even the specific telepresence application that we opened the thesis with was already imagined 20 years ago [121]. One early implementation of AAR was the augmented reality audio (ARA) mixer [122], which was based on an earpiece that allowed for active hear-through with built-in microphones, loudspeakers, and analog circuitry to realize required filtering. Work on a digital implementation of such

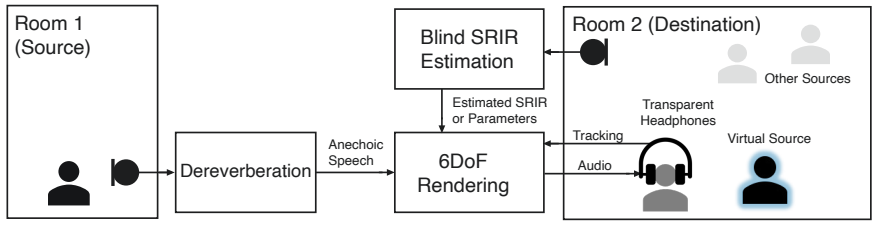


Figure 3.1. An example of an AR telepresence system. The participant in Room 1 is reproduced as a virtual sound source for the participant with headphones in Room 2. Other sources are present in Room 2.

(a) **(b)** **(c)**

Figure 3.2. Audiovisual speech recording. (a) Recording setup in the anechoic chamber "Lampio", incl. greenscreen, cameras, lights, and microphones. (b) The author infront of the greenscreen. (c) The rendered 3D point cloud video of the author in a room model of "Arni" — the visual equivalent of the virtual sound source in Fig. 3.1.

Figure 3.3. Real room and room model of the variable acoustics room "Arni".

active hear-through was presented, for example, in [123]. [124] gives a recent overview of the technologies involved, as well as applications, and challenges of such devices, which are now often referred to as *hearables*. In such earlier developments of AR audio, the emphasis was on these active hear-through technologies. Many of the more recent studies regarding perceptual requirements for AR, however, decided to use passively transparent headphones and focus more on other perceptual issues, such as room acoustic match, as Publication VI.

As seen in Fig. 3.1, a source signal for an AR telepresence rendering system should be anechoic, just as in any measurement-based virtual acoustic system. Residual reverberation on the source signal would lead to a *room-in-room* response [125, 126], which could potentially impair rendering. An anechoic signal might be obtained by recording it close to the user's mouth. In addition, enhancement methods based, for example, on near-field beam-forming or dereverberation [127] might be used to remove residual room information. In a full duplex system, with identification and rendering on both sides, having access to a dereverberated signal may even aid blind identification, as we have argued in [128].

Yet, dereverberation techniques are not part of this thesis. For experiments without real-time communication between parties, we can rely on recordings that are already anechoic, produced in the anechoic chamber. For this purpose, a new dataset of anechoic speech with 21 participants has been recorded. Participants were asked to read two sets of Harvard sentences [129], as well as parts of scripted conversations. The speech recordings were used for Publication VI. In addition to the audio data, also 3D point cloud videos were made, see Fig. 3.2c. These videos can be used for visual AR rendering using an HMD, which was done in [130] for a study regarding audiovisual congruence. One could see the process outlined in Fig 3.2 as the visual equivalent of AR audio rendering. Here, a visual reproduction of the speaker is captured that it then placed into the real environment of the user. Since no AR technology with good visual passthrough for showing the real environment was available at the time, in [130], the videos were shown in a model of the room, aligned to the real space. Fig. 3.3 shows a part of that model. This room model also enabled the experiments described in Publication VII, where not only the aligned room was shown, but also misaligned versions of it.

3.1 Blind Room Estimation

Currently, matching the virtual room acoustics to a given real room is seen as one of the important problems for the seamless integration of virtual sound sources into the real environment [131, 132]. In theory, this could be achieved using room acoustic simulations of the user's environment. However, that would presuppose sufficient knowledge of the room geometry and all surface properties. It may be possible to obtain this information using computer vision techniques, which was explored, for example, in [133–136]. Yet, currently, it appears challenging to infer acoustic properties of different absorbing materials purely from vision, so it is unclear if the accuracy of such techniques can be high enough to create auditory illusions.

A potentially more promising solution is to estimate SRIRs directly from acoustic input. In estimation-based systems, sound in the room of the user is used to gain an estimate of the room impulse response or of some acoustic parameters that can be utilized for rendering in a parametric reverberation algorithm. In principle, any signal present in the room may be used to perform such estimation.

Various deep-learning methods have recently been employed to perform single-channel RIR estimation. In general, they feature an encoder that extracts room acoustic information from the recorded signals and a generator that uses this information to synthesize a RIR. One of the first RIR estimation models with this structure was presented in [137]. Recent improvements using different models are [138–140]. Also, [141] follows

the same general architecture but employs efficient reverberators like feedback delay networks (FDNs) or filtered velvet noise at the synthesis stage so that rendering is computationally more efficient for real-time systems. Another related, early approach is the signal-to-signal room acoustic matching task in [142].

In most cases, a single, monaural speech source is assumed as the input. For AR practice, however, real scenes from which information can be extracted are likely to be more complex, featuring several different sound sources. So far, only [143] deals with multiple sources. It uses the same architecture as [137] but is trained to predict one RIR at a given distance, independent of the number of sound sources.

Special cases of the general estimator/generator architecture are systems in which one or both of the stages are not data-based, or the two are not trained together. The estimator may, for example, be designed to output a set of well-interpretable room acoustic parameters such as reverberation time (RT) and direct-to-reverberant energy ratio (DRR). An overview of such parameter estimation techniques based on digital signal processing (DSP) approaches for such parameter estimation is offered by the results of the ACE challenge [144], conducted in 2016. Since then, more and more deep learning methods have been employed for blind parameter estimation, see for example [145, 146].

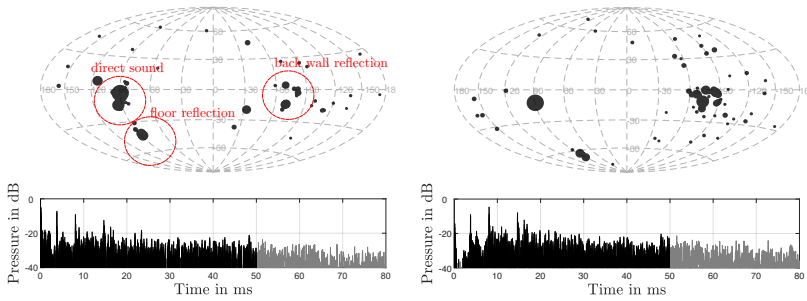
Deep learning methods for complete RIR estimation, or for parameter estimation provide promising directions for future work. However, classical DSP approaches should not be neglected completely, for they may provide robust and computationally inexpensive solutions that are more easily applicable in practice. In the past, DSP-based blind RIR estimation has often used cross-relation methods employing multiple distributed microphones [147–149], or adaptive filtering techniques, like LMS or frequency domain LS [150] that identify the response between a close reference and a more reverberant target. The problem with the latter is to find a suitable reference signal.

Estimation using a Pseudo-Reference Signal In Publication III, a simple DSP-based SRIR estimation approach was proposed. It is based on creating a *pseudo-reference signal* (a term not used in Publication III yet, but coined in [151] and adopted in [128]).

For the included publication, the pseudo-reference was obtained by beamforming. For this, a speech signal was recorded using a spherical microphone array, and the output signal was encoded to the SH domain. Then, an axis-symmetric, SH-domain beamformer, see Eq. (3.1), was applied to the SH-domain mic array signal as in

$$x_\mathrm{s}(\omega) = \boldsymbol{d}^\mathrm{T}\breve{\boldsymbol{x}}(\omega), \tag{3.1}$$

to obtain an estimate of the source signal x_s as a pseudo-reference signal.



(a) Measurement. **(b)** Estimate from speech input signal.

Figure 3.4. Comparison of pressure and DoAs between measured and estimated SRIR in the early part. The source was placed on the left of the array. A first-order floor reflection and back wall reflection are marked in (a), which are clearly identified in (b) — the output of the algorithm proposed in Publication III.

The next step can be thought of as *deconvolving* the Ambisonics signal with the beamformer signal, i.e., dividing them in the frequency domain

$$\check{\boldsymbol{v}}(\omega) = \frac{\check{\boldsymbol{x}}(\omega)}{\boldsymbol{d}^\top \check{\boldsymbol{x}}(\omega)} = \frac{\check{\boldsymbol{h}}(\omega)s(\omega)}{\boldsymbol{d}^\top \check{\boldsymbol{h}}(\omega)s(\omega)} = \frac{\check{\boldsymbol{h}}(\omega)}{\boldsymbol{d}^\top \check{\boldsymbol{h}}(\omega)}, \tag{3.2}$$

where we refer to $\check{\boldsymbol{v}}(\omega)$ as a relative transfer-function. Note that in [152], a similar concept is called "relative harmonics coefficient", and it can be seen as a special case of the "generalized frequency domain velocity vector" in [153], too. Having plugged in the SRIR convolved with the source signal $\check{\boldsymbol{h}}(\omega)s(\omega)$ in the above, it becomes clear that if the beamformer were able to extract the direct sound only, i.e., $\boldsymbol{d}^\top \check{\boldsymbol{h}}(\omega) = 1$, then $\check{\boldsymbol{v}}$ would be equal to $\check{\boldsymbol{h}}$. In practice, however, the beamformer output will also contain some room information. $\boldsymbol{d}^\top \check{\boldsymbol{h}}(\omega)$ might not even be minimum-phase as a consequence and the inverse $(\boldsymbol{d}^\top \check{\boldsymbol{h}}(\omega))^{-1}$ might not be stable and causal. In Publication V, these issues were not addressed explicitly. Rather, the acausal part of the estimated SRIR is simply removed. In addition, it was found that if the order of the beamformer is sufficiently high, the acausal part has low energy compared to the causal part and the identified response approximates the true response reasonably well in terms of reverberation time, and spatial features. Recently, [153] has extended the approach to retrieve estimates of $\check{\boldsymbol{h}}$ from $\check{\boldsymbol{v}}$ even in cases where impairments through a non-ideal reference are present.

What is considered in Publication V is that performing direct, instantaneous deconvolution of a block of signal as above does not give stable results in practice. Therefore frequency-domain, recursive estimation was applied to estimate the response between the pseudo-reference signal and all other SH channels.

For each frequency ω, the update equations are

$$\check{\boldsymbol{v}}(k) = \check{\boldsymbol{v}}(k-1) + \frac{1}{\phi(k)} x_{\mathrm{s}}^{*}(k)\varepsilon(k) \tag{3.3}$$

$$\phi(k) = \lambda\phi(k-1) + x_{\mathrm{s}}^{*}(k)x_{\mathrm{s}}(k) \tag{3.4}$$

$$\epsilon(k) = \check{\boldsymbol{x}}(k) - \check{\boldsymbol{v}}(k-1)x_{\mathrm{s}}(k), \tag{3.5}$$

where k is the index of a block of signal of a given length L that should be at least as long as the estimated response. An estimated response is compared to a measured response in Fig. 3.4.

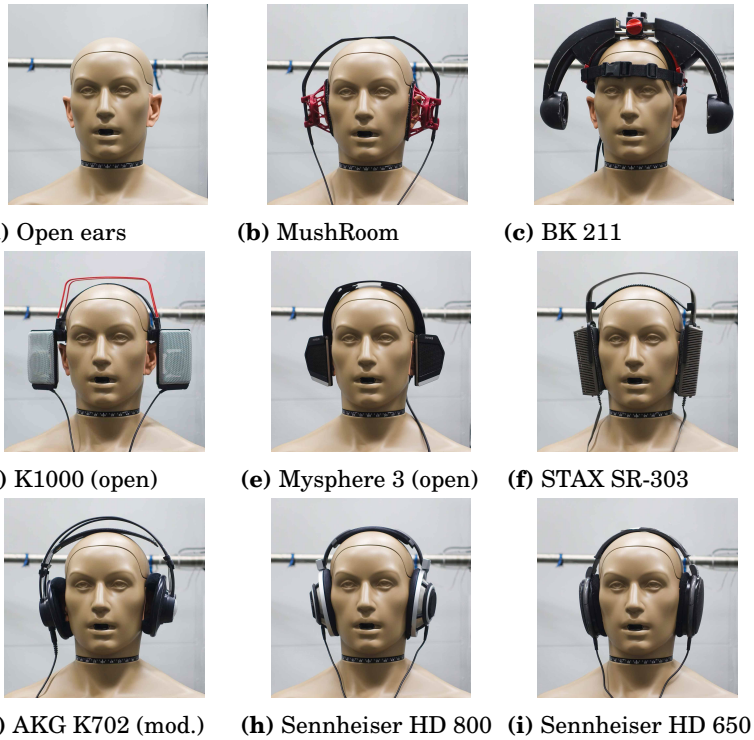
In addition to the recent developments by [153], in [128] the method was improved upon by including dereverberation to remove more room information from the pseudo-reference, thereby improving estimation. Also, we report estimation based on smart glasses rather than a spherical microphone array.

3.2 Acoustically Transparent Headphones

Another AR-specific question regards the used headphones. Interestingly, in earlier AAR research, active hear-through was the main focus [122–124], whereas more recently, passively transparent open headphones have been the more common choice for research regarding sound in AR. These headphones achieve transparency only through their open design. Fig. 3.5 shows the most common headphone models employed in AR tests. The AKG K1000, for example, was used in several AR-related experiments [154–156], as was the Mysphere 3 [36, 107, 117]. STAX electrostatic headphones were employed in plausibility tests [157, 158], although different models were used there. The modified AKG K702 is an inexpensive DIY solution [159] that was used in the so-called augmented practice room project [160], and in Publication V. The BK211 was a custom design [161], used for example in the authenticity test in [8]. Sennheiser HD650 and HD800 are mid- and high-quality regular, open headphones. The MushRoom headphone is a recent design [162, 163] that was employed in Publication VI.

The reason for choosing passive solutions over active hear-through for certain AR experiments is to avoid active DSP influencing real-world sound sources. However, also passively transparent headphones cause some perceptual impairments to real world sound. Since the real condition in real/virtual tests always depends on these impairments, they need to be thoroughly assessed, which was done in Publication IV.

Objectively, the transparency can be described through head-related transfer function measurements of subjects wearing the device and relating these measurements to the HRTF of the listener without wearing it. A simple ratio of the transfer functions, however, does not reveal any of the



(a) Open ears **(b)** MushRoom **(c)** BK 211

(d) K1000 (open) **(e)** Mysphere 3 (open) **(f)** STAX SR-303

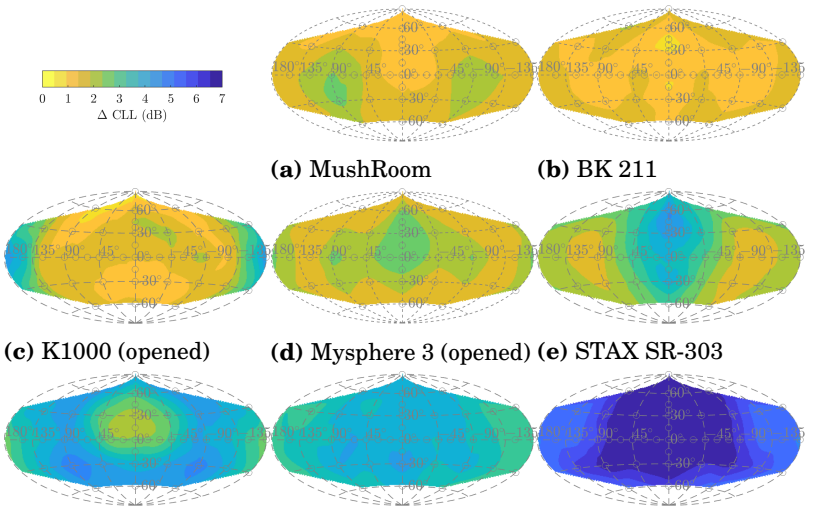
(g) AKG K702 (mod.) **(h)** Sennheiser HD 800 **(i)** Sennheiser HD 650

Figure 3.5. All tested models on the KEMAR HATS. (Pictures by Alexander Mülleder)

perceptual consequences. The most important effects on real-world sound when wearing headphones are changes in coloration and impairments of localization ability. In Publication IV, auditory models were employed for predicting these two perceptual implications.

Coloration Impairments in coloration can be modeled using a very simple auditory model. In Publication IV, we employed the composite loudness level (CLL). Therein, the model predictions were correlated with perceptual results from a listening experiment. A large difference in CLL implies a large perceived coloration. Based on these results, maps of CLL differences can be created, illustrating how much coloration of real-world sounds is expected when wearing a pair of headphones. Note that in Publication IV, evaluation was based on five head-worn devices and one set of measurements with a KEMAR HATS. During a later measurement campaign, the models depicted in Fig. 3.5 were measured with the help of eight participants replacing each headphone 15 times.[1] Fig. 3.6 shows CLL maps for these headphones. For the figure, the mean of the CLL values for eight participants and all replacements was computed. The least coloration is predicted for the special, extra-aural BK211.

[1] A publication with more detailed analysis of this data is in progress.



(a) MushRoom **(b)** BK 211

(c) K1000 (opened) **(d)** Mysphere 3 (opened) **(e)** STAX SR-303

(f) AKG K702 (modified) **(g)** Sennheiser HD 800 **(h)** Sennheiser HD 650

Figure 3.6. Model-based evaluation of coloration of all headphones, based on measurements on eight participants.

Yet the BK211 is also special in terms of its weight and form factor. The MushRoom headphone, developed in [162] appears to have the second lowest coloration, which is mainly concentrated in the lateral directions. As already observed in Publication IV, extra-aural like the K1000 or the Mysphere 3 impose much less coloration than traditional, "open" headphones like the Sennheiser HD650.

Localization Impairments Localization impairments can also be predicted using auditory models. The largest effects were shown for vertical localization; horizontal localization ability is usually affected to a lesser extent, as shown in Publication IV and in [164], where the influence of wearing STAX headphones on horizontal localization was on average 0.6°. For evaluation of vertical localization Publication IV proposed to use the sagittal plane localization model by Baumgartner [165], as implemented in the Auditory Modelling Toolbox [166]. It compares a set of HRTF templates with a set of target HRTFs. Here, the templates were regular, open-ear HRTFs and the target were HRTFs measured while wearing the headphones. The model relies on positive spectral gradients as features and outputs probabilistic estimates of sagittal plane localization. Fig. 3.7 shows predicted quadrant errors based on these outputs for the tested headphones, using data from eight participants, replacing each headphone 15 times each. Like this, distributions can be compared rather than single values as in Publication VI. BK211, K1000, and MushRoom show the lowest impairments.

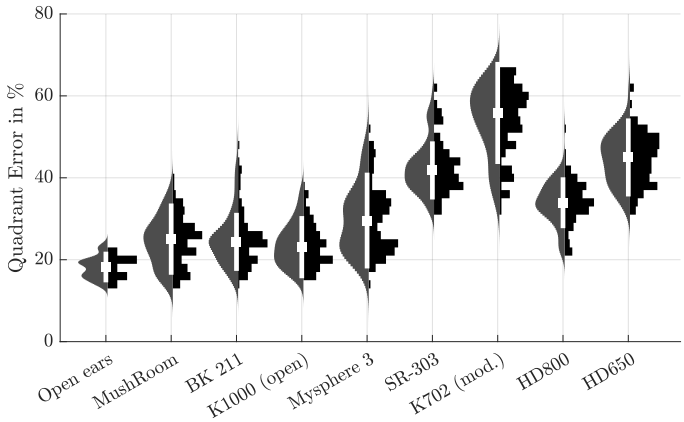


Figure 3.7. Predicted quadrant error when wearing different headphones. Histograms on the right in black, and violin plots on the left in gray.

4. Introducing Evaluation Paradigms for Sound in AR

Now that contributions to virtual acoustics technology in general and techniques specific to AR have been presented, evaluation methods for sound in AR are discussed. In the history of virtual acoustics, rendering was often evaluated in "virtual only" tests by direct comparison to a high-quality, yet virtual, reference. In such tests, participants are asked to rate similarity, overall quality, or specific auditory attributes of the rendering in relation to the reference. The International Telecommunication Union (ITU) provides several recommendations for running such tests. ITU-R BS.1116-3 [167] defines a method that is based on direct comparison of a stimulus to an open and a hidden reference; ITU-R BS.1534-3 [168] defines the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) method, where several stimuli are compared to each other, as well as to a reference and to anchors. MUSHRA or MUSHRA-like designs are commonly employed, also in the context of AR/VR [169]. A more detailed discussion of these methods, as well as other evaluation methods used for AR/VR can be found in [170]. Often, the results of tests involving direct comparisons to a reference indicate that virtual acoustic rendering quality is very good, but that some differences to the reference remain; see [14, 22, 171–174] for results using different room acoustic simulation techniques from different decades, and [77, 80, 100, 104, 175] for techniques based on microphone array measurements.

Fortunately, situations exist in which audible differences are tolerable; the specific requirements always depend on the application. Publication VII shows a good example. It demonstrates that clearly audible, position-dependent room acoustic differences oftentimes do not help subjects localize themselves in a room. Listening to the acoustics from one zone within a room does not keep participants from firmly believing that they are positioned in a completely different zone, given that the direct sound remains the same. This result in itself matters more for VR than for AR, where it could mean that not too much effort should be spent on position-dependent room acoustic differences. In any case, it highlights that the importance of room acoustic differences is task- and application-dependent.

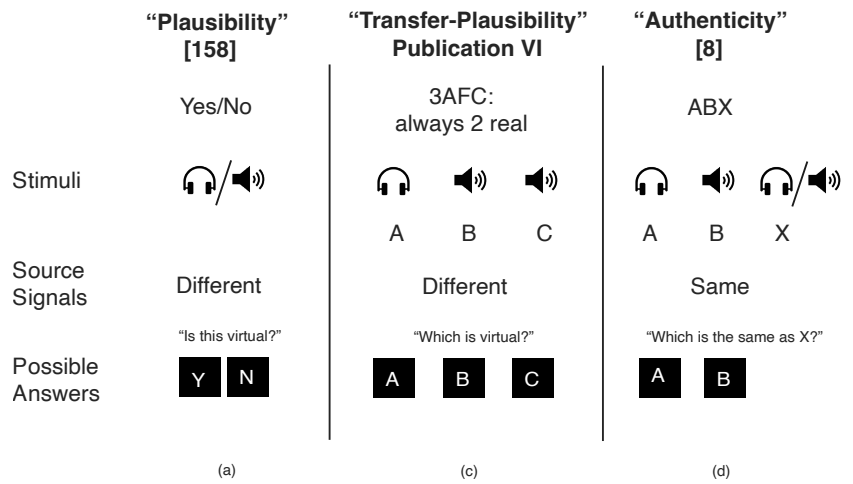


Figure 4.1. Overview of paradigms for evaluating sound in AR based on the emergence of auditory illusions. In the Yes/No plausibility test, only one stimulus is presented, either real or virtual, and the subject must decide whether it is virtual. In the 3AFC transfer-plausibility experiment, the task is to detect the virtual source amongst three sources using different signals. There are always two real sources and one virtual source. In authenticity, all renderings are presented using the same signals from the same spatial locations, and subjects need to decide if "A" or "B" is the same as the reference ,"X", which can be either real or virtual.

Consequently, sound for AR should be tested using experimental paradigms specifically tailored to AR, matching its special requirements. The first idea that suggests itself when thinking about possible tests for AR is to use paradigms involving both real and virtual sound sources. The earliest examples of tests that compare real and binaural virtual sound are [176–178]. They all tested anechoic sources under various paradigms.

Amongst all paradigms involving real and virtual sources, we distinguish between those that test whether an *auditory illusion* [7] is evoked and those that do not. In this context, an auditory illusion is said to take place if a participant believes a virtual sound source to be real. In previous literature, authenticity [8] and plausibility [158] have been proposed as two possible paradigms that indicate such illusions. In the following, we discuss these concepts, along with the notion of transfer-plausibility. Fig. 4.1 gives an overview of these three paradigms and their associated experimental designs. Publication VI compares them in a test regarding mismatched room acoustics.

Moreover, this chapter proposes a conceptual model encompassing the different paradigms, technical and perceptual factors mentioned so far. Instead of classifying paradigms by their reliance on inner and external reference, it distinguishes them by the degree to which they depend on

auditory similarity, *artifacts*, and *context*. At the end of the chapter, we offer opportunities for future AR audio experiments, for which groundwork has been laid throughout this work.

4.1 Auditory Illusions and the Edison Test

The goal of reproducing sound to evoke the illusion of being real is immanent in audio technology. Between 1915 and 1925, thousands of so-called “tone tests" were conducted to promote the new, diamond disk phonograph by the Edison company [179]. In these tests, listeners compared sound reproduced using the device to performances of real musicians. Consequently, the comparison was between a loudspeaker reproduction and the real physical sources. In the AR tests conducted today, the virtual condition is a headphone rendering, and the real condition is usually a loudspeaker reproduction already. The reason for using loudspeakers as real sources is the increased reproducibility and controllability compared to using other physical sound sources. However, the paradigms introduced in this chapter could, in principle, be applied to comparisons between headphone rendering and any kind of physical sound source, too.

What is surprising about the historical “tone-tests” is that several newspaper reviews from the time reported that correctly identifying the real source was difficult. In retrospect, it is unclear whether these reviews were overly enthusiastic, or whether listening to recorded sound was still so unfamiliar to many people that they were not sensitive to noise and limited bandwidth, which is expected from a phonograph of that time. Unfortunately, no response data for specific tests are available today. After all, the purpose of these tests was advertisement, not scientific research. Given this purpose, it is not surprising that when looking at specific programs that are reported, it appears that a mix of different kinds of performances was presented, under no strict experimental paradigm. For example, musicians played in ensembles, with recorded accompaniment, or singers sang duets with recordings of themselves, sometimes simultaneously and sometimes alternating between real performance and recording between phrases [179]. It is also reported that a musician left the stage in the dark [179], with the machine creating the illusion of a real musician on stage. Interestingly, these different scenarios correspond rather closely to different versions of the paradigms that test for auditory illusions shown in Fig. 4.1 and discussed in more detail below.

4.2 Tests not requiring Auditory Illusions

Even experimental designs that include both real and virtual sound sources do not necessarily require auditory illusions to take place. Examples of this are all designs in which participants rate one or more virtual conditions compared to a real reference, regarding different auditory attributes. Attributes could stem from the spatial audio quality inventory (SAQI) [180], the inventory shown in [181], or the recent immersive music experience inventory [182], for example. In [183,184], participants rated virtual sound sources in reference to a real sound source concerning the attributes "reverberance", "source width", "source distance", "source direction", and "overall quality". Such attribute ratings can give insights into specific differences. One disadvantage, however, is that participants need to be skilled enough to interpret the attributes correctly, whereas an auditory illusion test can be done by anyone.

Other tests that do not require auditory illusions per se use "plausibility" as an attribute, rated on a scale [155, 185–187]. This implies an understanding of plausibility as a *matter of degree*, which allows for gradual differences in rating. The thorough discussion of the term by Kuhn-Rahloff adopts this view [188, p. 42]. Following from his considerations and experiments, he defines plausibility in the following way:

> Plausibility is the result of a perceptual process determining to which extent perceptual objects (German: Wahrnehmungsobjekte) agree to an inner reference, which depends on individual previous experience. [...]

This definition of plausibility is wider than those we discuss in the next section; it also encompasses cases, in which no real counterpart to the simulation exists. Such cases are more likely to occur in VR than AR.

Another important type of experiment that does not directly require auditory illusions tests for externalization, i.e., the perception that a sound source is outside of the head [189]. The most common experimental design testing for externalization lets participants indicate the position of a sound source on a continuous scale reaching from inside the head to a fixed distance in external space [189]. In our view, externalization is a prerequisite for creating an auditory illusion, i.e., a sound that is perceived inside the head will not be perceived as real. However, we have observed that cases exist in which sound is well externalized, but still recognizable as virtual. For example, in [107], where participants were asked to indicate reasons for not perceiving a source as real, externalization was only the fourth most commonly mentioned issue. Externalization was the main outcome variable in experiments regarding the so-called room divergence effect, where it was shown to be reduced when there was a mismatch between the acoustics of a real room and a reproduction [190]. Next, we discuss paradigms that test whether an auditory illusion took place or not.

4.3 Plausibility

Lindau and Weinzierl define a plausible rendering as

> a simulation in agreement with the listener's expectation towards an equivalent real acoustic event.

This wide definition is similar that of Kuhn-Rahloff, but replaces the notion of the "inner reference" with that of "expectation" and excludes "perceptual objects" that do not occur in the real world. It can serve as a general, overarching theme for several experiments discussed in the following sections. The experimental design proposed in [158], provides a specific operationalization of plausibility. Therein, the question is whether a single sound source presented in isolation is believed to be real, i.e., whether it evokes an auditory illusion. To test this, a single stimulus is presented at the time and participants need to decide if it is real or virtual, see the left column of Fig. 4.1. Note that this presupposes a completely different interpretation of the term as in Kuhn-Rahloff, not as a matter of degree, but as a binary variable: either an illusion is created in a particular trial, or it is not. Plausibility tests in this sense were conducted in [156–158, 191, 192].

In the light of the operationalization as a Yes/No test, another understanding of a "plausible" system is implied.

> A rendering system is plausible if virtual sound sources are *believed to be real* when presented in isolation.

Note that when we speak of "plausibility in the sense of [158] or "plausibility in a yes/no task" in the following sections and in Publication V and VI, it is the precise experimental design and this understanding that we refer to. In experiments in this sense, listeners should rely only on their expectation of a sound in the real world when listening to the rendering, without the possibility of comparing it to a real reference. Thus the term *inner reference* as discussed by [188] is used here as well.

A possible risk of plausibility experiments in this particular sense is that this inner reference could be variable throughout the test so that subsequent trials influence each other. For example, [156] has shown that leaving real trials out of the experiment influences the results. This points to the fact that comparisons rarely occur only to an inner reference but also to other stimuli presented in the test, which is especially problematic when different rendering conditions are to be compared in one test — a common scenario when developing new systems. The term "plausibility with a tuned reference" [156] has been used when other sound sources may influence the outcome. Such sequential effects are attempted to be reduced

by using different signals and different spatial locations in subsequent trials. However, it is unlikely that they can be removed completely.

Regarding data analysis, a Yes/No task requires taking individual bias into account. This is usually done by signal detection theory (SDT) analysis, where the outcome measure is the sensitivity, d'. SDT analysis is described in more detail in [158] and Publication VI.

4.4 Authenticity

Picturing the plausibility paradigm in the sense of [158] on one end of a scale, where no comparison to an external reference should be possible, the *authenticity* paradigm would be on the other end. There, the aim is to test if a real source and a virtual version thereof cannot be distinguished at all. We could define

> A virtual source is authentic if it is indistinguishable from its real reference.

Blauert [193, p. 358] introduces authenticity in the following way: "If the reproduced signals in the ear canal correspond exactly to those during recording, the spatial attributes of the auditory events also correspond exactly to those that the subject had during the recording. An 'authentic' spatial reproduction thus takes place".

Taking a closer look reveals two concepts, which could be framed as "physical authenticity" and "perceptual authenticity". The quote could be understood as saying that if physical authenticity is achieved, i.e., the pressure at the eardrum is reproduced correctly, "perceptual authenticity" follows. When speaking of authenticity in the following, we always refer to "perceptual authenticity", as exact reproduction will never be achieved up to arbitrary precision. Some physical differences will always persist, may they only be due to time-dependent variation of the room acoustics, whereas perceptual indistinguishability is at least theoretically possible.

To assess perceptual authenticity in the sense of indistinguishability, discrimination experiments between real sound sources and virtual renderings need to be performed. In the case of authenticity, subjects are not asked to decide whether renderings are real or virtual, but the test is designed to check for any perceivable difference between them. Any audible difference would contradict indistinguishability and, thereby also, authenticity. A suitable experimental design is the ABX test as used for the authenticity experiment regarding virtual acoustic rendering in rooms by [8]. The outcome variable simply is the proportion of correct responses.

Authenticity is an extremely strict requirement for evaluating AR systems that is difficult to meet in practice. As discussed in several places throughout this thesis, indistinguishability between a rendering and a reference is rare in virtual acoustics, even if the reference is only a high-

quality reference rendering; using a real source cannot be expected to make comparison less sensitive. Fortunately, direct comparisons between a real source and a virtual version thereof are usually impossible in practical AR applications. Note that even if authenticity experiments might not be the most relevant choice for evaluating practical AR systems, there exist cases in which they are highly valuable. Also, if it was possible to show that a virtual acoustic system is authentic, it could be used as a reference instead of the real world in future developments of practical AR systems.

4.5 Transfer-Plausibility

In a practical AR telepresence application, it can be expected that different real sound sources are active alongside virtual sources, emitting different signals from different spatial locations. This allows for a certain degree of comparability between real and virtual sound. As opposed to authenticity tests, this comparison is not direct. Instead, listeners need to cognitively *transfer* properties between real and virtual sources. In the case of room acoustic differences, for example, listeners first need to separate the source signal from the room acoustics, which has been shown to be possible through the statistical regularity of reverberation [194]. Still, in [195] we have shown that differences between room acoustical conditions become less noticeable when different speech signals are used. The same was found when different musical excerpts were used for comparing concert hall acoustics [196].

To arrive at an experimental paradigm, we define

> A virtual sound source is transfer-plausible if it is *believed to be real* in the presence of real sound sources.

This is operationalized by a test in which at least three *different* sound sources are presented, one of which is virtual. Participants then need to decide which one of them is the virtual source. With three sources, this is a 3-alternative-forced-choice (3AFC) task. Out of the three sources, two are always real, see Fig. 4.1. If only one real and one virtual sound source were used, participants may detect a relevant difference, but not know which of the two versions belongs to the real world. Even though it is conceptually different, such a 2AFC experiment can be conducted as well. It is included in Publication VI, where the results are compared to the Yes/No (Plausibility in the sense of [158]) and the 3AFC (Transfer-Plausibility) test just described.

These choices and definitions were not as clear when first introducing the idea of transfer-plausibility in Publication V. Therein, a simultaneous detection and identification task was performed. That is, there were also

renderings without virtual sources at all. While this is a possible, and interesting experiment as well, it makes interpretation of the results more difficult. Decision models would need to be adapted from lineups performed in eyewitness testimonials, e.g. [197].

One could say that transfer-plausibility makes the comparison to a real-world sound explicit, which is implicit and hard to control in plausibility tests. In TP, it is clear that other, real sound sources are available to the listener. Apart from this, an advantage compared to Yes/No tasks is that transfer-plausibility tests are AFC designs rather than Yes/No tasks so that they do not depend on individual bias; the proportion of correct identifications of the virtual source can simply be averaged across participants, as in the ABX authenticity test, without having to compute d' scores for each participant. Especially in tests comparing different rendering conditions, there may be very few answers for each condition and each participant to do so. d' estimates from a low number of trials are seen as problematic, an issue that is discussed in Publication VI. What all paradigms share is that changes in performance might occur during the experiment. Also, while systems can be compared within and between tests, an absolute threshold at which a system could be called "plausible" or "transfer-plausible" will always be an arbitrary definition.

The concept of transfer-plausibility is similar to what has been introduced as *co-immersion* by [198, 199]. However, therein no experiments with real and virtual sound sources have been conducted. The real condition corresponded to a high-quality rendering rather than a sound source in a room. In [195, 200], we called a similar test "transferring task"; properties of different sources are transferred, but not between real and virtual sources.

4.6 Comparing Paradigms

The experiments presented in Publication VI illuminate the influence of room acoustic match on the emergence of auditory illusions. For the test, a simple, static rendering system was used, in order to exclude effects of tracking or impairments caused by interpolation. Four experimental designs were implemented in Publication VI: authenticity (in an ABX test using the same source signals), plausibility (in a Yes/No task), transfer-plausibility (in a 3AFC task, where always two sources were real), and a 2AFC task (with one real and one virtual source).

The main result of the study is that transfer-plausibility, as a new experimental paradigm, can resolve differences between non-ideal rendering conditions. In contrast, in the authenticity test, all non-ideal rendering conditions featuring room acoustic mismatch were detected with almost perfect accuracy. This shows that the authenticity paradigm is not well

suited for comparing different non-ideal rendering conditions, which is commonly of interest when developing AR systems.

Moreover, the test studied differences between the 3AFC transfer-plausibility test and the 2AFC task, in which one real and one virtual source were presented in each trial. After considering the different guessing rates, participants were more sensitive in the 3AFC transfer-plausibility paradigm than in the 2AFC test. This highlights that the presence of a second, real source provides participants with more information.

Comparing the results between tests also revealed that the Yes/No plausibility test led to higher sensitivity values than expected from the decision theory when compared to the 2AFC test. We attribute this to the fact that some comparison is indeed possible across trials in a Yes/No task. A similar result can be observed for the comparison of Yes/No task, and 2AFC task in [178], where anechoic real and virtual sound sources were rendered using the same source signals. Therein, the percentage correct for both paradigms is nearly identical, whereas detection theory would predict a difference.

4.7 The Auditory Similarity / Artifacts / Context Model

In Publication V and VI, it was shown that different experiments involving auditory illusions lead to different results, and one may naturally ask why this is the case. In the following section, a hypothetical model is proposed to explain some of these differences. It should aid in relating other experiments that do not involve creating illusions to the discussed tests and clarify the influence of certain percepts. We suggest the model to encompass three aspects: *auditory similarity, context* and *artifacts*.

Therein, *auditory similarity* depends on auditory sensory memory (ASM). ASM is a construct used to explain the first stage of working memory, and is especially important in memory for non-verbal qualities of sounds [203–205]. ASM is expected to decay within a few seconds. During this time, a listener retains access to a heard signal, and information can be extracted from it, which is encoded into working memory. If another signal is presented during this time span, much smaller differences can be detected between the two than after longer retention intervals. After longer times, only high-level, verbal descriptors that were extracted from the sound are left in memory. There are good arguments for the existence of ASM. First and foremost, almost all sounds have some temporal extent. To recognize a unit of sound, e.g., a phoneme, the signal needs to be "buffered" until the unit can be processed and meaning can be extracted. Nevertheless, the memory of non-verbal sounds is studied far less than the memory of verbal sounds [206]. Particularly, research on the memory for

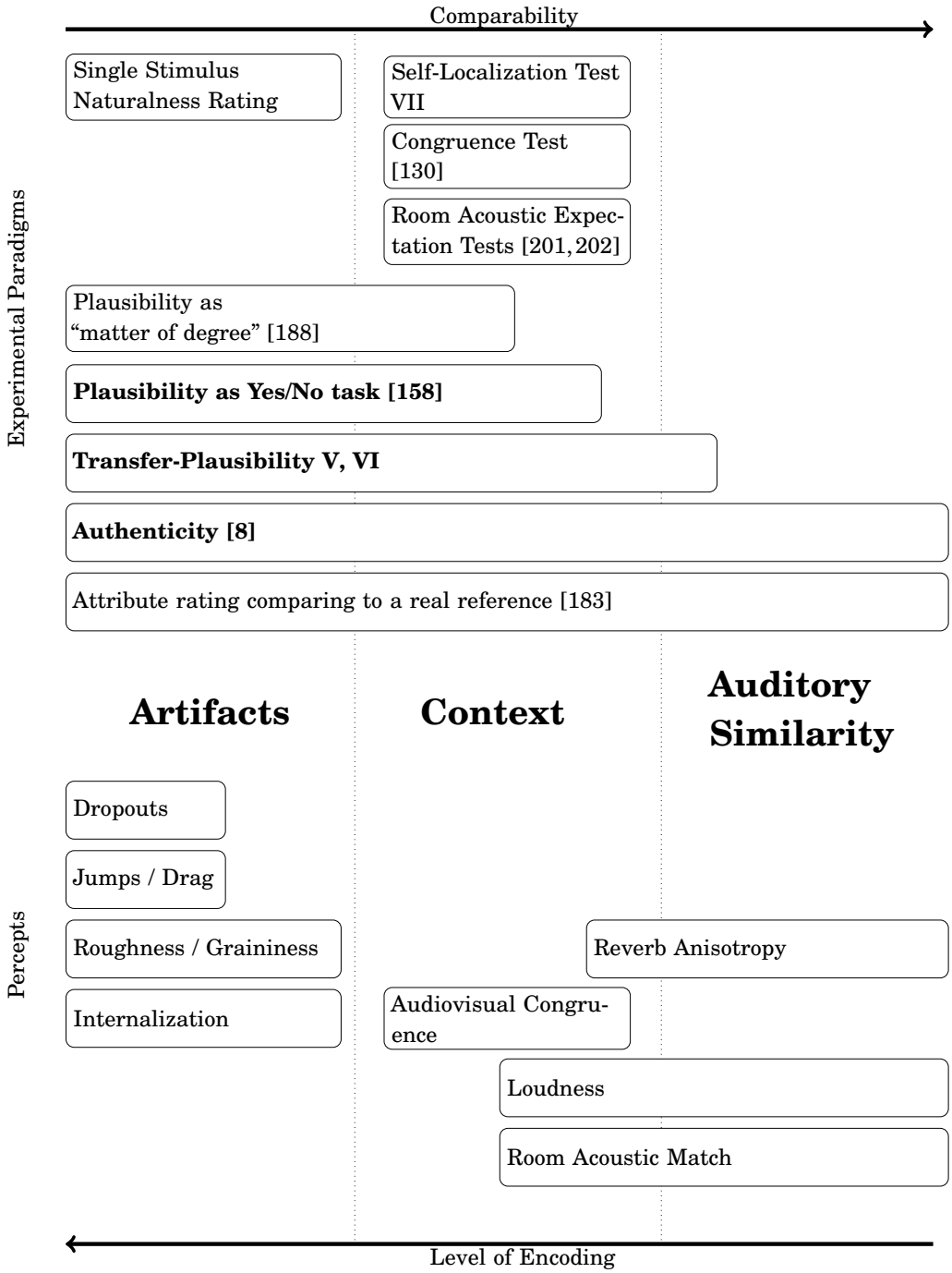


Figure 4.2. Experimental designs and certain percepts and their hypothetical mapping onto the similarity / context / artifact model. Experiments testing for auditory illusions are highlighted.

room acoustics, which can be a deciding factor in the emergence of auditory illusions, has only recently begun [207, 208].

Context includes aspects of a sound that are available past this sensory memory decay , either through cross-modal input (typically visual) or through accessing long-term memory of other sounds encoded from sensory memory at earlier times.

Finally, *artifacts* refer to auditory aspects that can be judged completely without context, other than all of a human's experience of the world, encoded over the lifespan.

This model may seem abstract at first, but we now concretize it by showing that different perceptual factors can be arranged according to their influence on the three model aspects and that different experiments are sensitive to the model aspects to various degrees, see Fig. 4.2. Special emphasis is put on aspects that relate to the studies in this thesis.

Tracking problems like jumps caused by loss of head tracking, or drag caused by excessive latency are typical examples of *artifacts*. They can be noticed completely independent of a specific context and do not depend on similarity to a reference. The same is true for audio dropouts, or specific artifacts of certain virtual acoustics rendering methods, like the roughness that can be caused by uncompensated SDM rendering of very transient sounds that Publication II deals with. Another example would be the "metalicness" that is sometimes associated with small, very ill-tuned FDNs, and is due to insufficiently dense and equal modal excitation [209]. Such aspects are noticed in reference-free tests. For example, by letting participants rate a single stimulus according to "naturalness", or "realness", which are attributes that refer to the real world as a whole.

Room acoustic mismatch, on the other hand, is not an artifact. Whether there is too much or too little perceived reverberation, cannot be judged in isolation. It can, however, easily be judged based on auditory similarity and, to some extent, by context; the required accuracy of acoustic match differs between experiments that depend on similarity and experiments that depend on context. As an example, the acoustics of two office rooms may be distinguishable in direct comparison, but both may be equally valid in an acoustic expectation test akin to [201, 202], in which a visual rendering of an office is shown and the question is if an acoustic rendering matches it. An auditory rendering of a cathedral, however, might not even be deemed fit in the expectation task, where a visual representation of an office is shown. The same is true for a slightly ill-adjusted reverberator. Even though it may be free of artifacts that are noticed when listening in isolation, and the rendering might fit a visual context, a difference may still be detected in terms of similarity to a reference. Loudness is another example of a percept that more strongly influences similarity experiments but also matters to some degree if context judgments take place. Subtle errors in rendering the spatial distribution of the late reverberation as

those analyzed in Publication I might mostly be detectable in tests that depend on similarity.

Furthermore, the self-localization test shown in Publication VII is an example of an experiment that only depends on context. There was no chance to check for auditory similarity, and even if there had been artifacts in the loudspeaker playback, they would have had no direct effect on the result. Tests for audiovisual congruence, e.g., [130] form another test class that only depends on context. If there were no visual cues in an audiovisual congruence test, it would be impossible to tell if a sound source originates from the correct direction.

More delicate differentiations need to be made when mapping the discussed auditory illusion tests onto the model. If a plausibility test is implemented as in [158], the goal appears to be that only context and artifacts are rated. However, as we argued in Publication VI, possible comparability across trials allows for auditory similarity that might have some influence on the result, making the test reach relatively far to the right in Fig. 4.2. In transfer-plausibility, auditory similarity plays a more explicit role, but since acoustic transferring is required, and audiovisual offsets might influence TP results, which can be seen as a kind of context. Clearly, there is no complete dependence on auditory similarity. In authenticity tests, any perceivable dissimilarity contradicts indistinguishability, so it depends strongly on auditory similarity. Artifacts influence all three paradigms equally.

It should be emphasized that the model just discussed is hypothetical so far. However, it is testable through future work by evaluating the same stimuli under different paradigms and quantifying the contributions of certain percepts, like errors and mismatches, to the outcome of all tests.

4.8 Outlook: Other Factors

Obviously, room acoustic mismatch studied in Publication VI is only one example of a factor that can be studied under the TP paradigm, and the influence of various other errors could be tested, too. Regarding some of these questions, we have laid the groundwork in this thesis and in other publications, but many aspects have not yet been tested in auditory illusion tests. The following aspects appear especially promising for future tests.

Scene complexity When the idea of transfer-plausibility was introduced in Publication V, it was tested in scenes of different complexity. In other words, different numbers of sound sources were presented simultaneously and the virtual source had to be selected amongst them. The percentage of correct answers was reduced when the number of sources was increased. At this earlier stage, the TP paradigm was not as well defined as above, and the fact that also the answer “there are no virtual sources” was pos-

sible made evaluation difficult. A future test could try to disentangle the effect of scene complexity as the number of simultaneous sources and the number of alternatives in the test by always conducting a 3AFC test as in Publication VI, while presenting different numbers of interfering sources. It is, however, worth discussing if very complex scenes are an important test case, regarding that many AR telepresence applications may involve only a few participants.

Position Dependency Publication VII indicated that self-localization from position-dependent acoustical differences is very difficult for untrained listeners. Publication VI showed that a response measured at a different position was slightly less (transfer-)plausible than the matched response. This difference should be evaluated further. A possible test could be a 6DoF test, where room acoustic rendering is either position-dependent, for example, by employing microphone measurements on a grid, or remains constant.

Objective Acoustic Parameters In Publication VI, the physical properties of the room were changed between rendering conditions, which had an influence on various room acoustic parameters. Instead of making physical changes, one measured response could also be manipulated, so that they only vary in certain objective parameters. Results in terms of reverberation time or direct-to-reverberant energy mismatches on TP, could help to interpret results from blind parameter estimation algorithms better. It should, however, be considered that not only are these parameters dependent in the sense that they both change when the physical room is changed, but they might also be perceptually dependent. In [200] we presented a 2D threshold test in a transferring paradigm, which could be converted to a transfer-plausibility test in the future.

Audiovisual Congruence For evoking auditory illusion, localization matters only in the sense that no incongruence between auditory and visual rendering should be noticed in order not to destroy the illusion. In [130], we described an audiovisual congruence test using speech stimuli, presenting either 3D point cloud videos of avatars or loudspeakers. Congruence was tested by moving the position of the visual rendering in relation to the sound source. For the avatar, the median of the offset at which 50% of renderings were perceived as incongruent was approximately 11° for horizontal offsets. The 50% point for detecting incongruency with vertical offsets was hardly reached in our test, where the maximal offset tested was about 20°. Assuming that localization errors only destroy auditory illusions if the audiovisual congruence range is exceeded, it seems that localization errors should not be a dominant problem, yet this should be tested explicitly.

Individual HRTFs Another obvious, related question is whether individual HRTFs are required for creating auditory illusions. Using non-individual HRTFs is known to increase vertical localization errors and front-back confusion rates [47]. Yet most auditory illusion tests conducted so far used non-individual HRTFs and some reached high confusion rates, e.g., Publication VI. This is likely because localization errors do not exceed the audiovisual congruence range, as explained above. Still, more front-back confusion could occur that breaks the illusion. They may partially be prevented in head-tracked systems, where front-back confusion caused by non-individual rendering is generally reduced [210, 211].

Headphones In Publication IV we have developed tools for assessing the perceptual transparency of headphones. What has not been tested yet is the influence of strongly impairing headphones on auditory illusions. Clearly, in practical applications, headphones should be as transparent as possible, as impairments to the real world are generally undesirable. Transfer-plausible rendering of virtual sources, however, could even be *easier* to achieve if both real and virtual sound sources are impaired. If, for example, localization accuracy is reduced for real sources, too, higher audiovisual mismatches for virtual sources may be tolerable when the aim is to evoke an illusion.

Moreover, in real-life implementations of AR telepresence systems, it can be expected that the used transducers may not deliver sufficient output at all audible frequencies. The influence of limited bandwidth would be straightforward to test in auditory illusion experiments as well.

Latency We introduced a simple method to measure the latency of 3DoF and 6DoF systems in [212]. The detectability of latency has been determined to be at approximately 50 - 60 ms [106]. If we take latency to belong to the artifact category, we would assume its detectability to depend relatively little on the paradigm, but TP tests could confirm this.

Reproduction Level The spectral content of speech changes depending on the vocal effort [213], amongst other features. Thus, human listeners might use it as a cue to form an expectation of absolute speech level. If the reproduction level is incorrect, which is likely to happen in practical systems, sources may tend to be perceived as virtual. In the future, mismatches in real and reproduced speech levels should be tested in auditory illusion tests.

Source Directivity So far, source directivity has not been mentioned at all, although direction-dependent spectral changes are often noticeable. As a whole, the directivity can have an influence on room acoustic perception. The directivity of human speech had recently been assessed rather extensively, e.g. in [214, 215], and TP tests could be conducted, which evaluate how precisely these need to be taken into account to create illusions.

Other system factors In addition to all these questions, the influence of specific technical choices in rendering systems could be tested directly. For example, [216] conducted a plausibility test for anechoic HRIR rendering based on Ambisonics of different orders. Obviously, the same could be done for checking the required order for SRIR, as it is done in [217] in a MUSHRA-like test. Also, different parametric SRIR rendering algorithms should be evaluated regarding auditory illusions as, for example, in [154], putting the results of Publication I and II into context. Ultimately, the goal of the field is to implement and evaluate complete systems, also involving blind estimation as in Publication III and efficient real-time rendering.

5. Summary and Conclusion

The beginning of this thesis described techniques for auralizing measured room acoustics. We have shown parametric SRIR processing as a method for creating 3DoF virtual acoustic rendering based on microphone array measurement. A particularly straightforward variant for this is SDM. Two fundamental properties and limitations of analysis and rendering stages in SDM were studied: Publication I showed that DoA estimation based on the pseudo intensity vector can capture anisotropy in the late part of a SRIR, but only to a certain extent for which an analytical model was derived. Since the resulting expression only has three degrees of freedom, complicated directional distributions are not detectable.

In SDM rendering, the spatial assignment can lead to roughness, which is audible when rendering very transient sounds. It was shown to be due to head shadowing in Publication II. Since then, methods for compensating roughness have been proposed [101, 102].

We have further argued that full 6DoF rendering will always need to apply sound field models similar to those employed in parametric SRIR processing methods where model-based SRIR processing has so far not led to results that were perceptually indistinguishable from a reference. Rendering is expected to be even less perfect in practical AR systems, where SRIRs need to be estimated rather than measured, by techniques such as the one proposed in Publication III. Therein, blind SRIR estimation based on forming a pseudo-reference signal based on beamforming was proposed. It was shown that if the order is high enough, certain temporal and spatial features can be estimated in this way. The approach was later extended in [128].

Publication IV demonstrates how real-world listening is impaired when wearing transparent headphones. Auditory models are introduced to predict the effect. They revealed that no headphones used in AR studies so far could be considered perfectly transparent but that the influence of the MushRoom headphones [162] used in Publication VI is relatively small, also when compared to the DIY model used in Publication V.

The expected limitations of parametric rendering, blind estimation, and headphone reproduction highlight that authenticity, i.e., the indistinguishability between real and virtual, is not an appropriate goal for AR. Publication VI demonstrated that cases exist, in which rendering is not authentic, but it can still cause an auditory illusion, i.e., a virtual sound source appears to be real. Different experimental paradigms that test for auditory illusions were discussed and tested in Publication V and VI. Apart from showing that (transfer-)plausible rendering is possible even in non-ideal conditions, it was demonstrated that plausibility does not solely rely on a so-called internal reference, or that the internal reference needs to be understood as constantly varying throughout the test. At the end of the last chapter, we have presented a hypothetical model for understanding the relationships between percepts and certain experiments by arranging them based on their contributions to three aspects: auditory similarity, artifacts, and context.

Finally, opportunities for future studies regarding the influence of certain perceptual factors on the emergence of auditory illusions were described. One question that has only been touched upon in Publication VI is the importance of correctly rendering acoustical differences that exist within a room. Publication VII has shown that at least, using such differences to localize oneself in a room is nearly impossible for untrained listeners.

References

[1] F. Denk, F. Brinkmann, A. Stirnemann, and B. Kollmeier, "The PIRATE: an anthropometric earPlug with exchangeable microphones for Individual Reliable Acquisition of Transfer functions at the Ear canal entrance," in *DAGA - Fortschritte der Akustik*, (Rostock, Germany), Mar. 2019.

[2] F. Zotter, "High-resolution directional impulse responses of the Eigenmike EM32," May 2018. https://phaidra.kug.ac.at/o:69292.

[3] S. H. Foster and E. M. Wenzel, "The Convolvotron: Real-time demonstration of reverberant virtual acoustic environments," *J. Acoust. Soc. Am.*, vol. 92, pp. 2376–2376, Oct. 1992.

[4] A. Neidhardt, F. Klein, N. Knoop, and T. Köllmer, "Flexible Python tool for dynamic binaural synthesis applications," in *142nd Audio Engineering Society Convention*, (Berlin, Germany), May 2017.

[5] L. McCormack and A. Politis, "SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods," in *AES International Conference on Immersive and Interactive Audio*, (York, UK), Mar. 2019.

[6] 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," in *24th International Congress on Acoustics*, (Gyeongju, Korea), Oct. 2022.

[7] K. Brandenburg, F. Klein, A. Neidhardt, U. Sloma, and S. Werner, "Creating Auditory Illusions with Binaural Technology," in *The Technology of Binaural Understanding*, Modern Acoustics and Signal Processing, Springer International Publishing, 2020.

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

[9] M. Vorländer, *Auralization: Fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. RWTHedition, Berlin: Springer, 1st ed., 2008.

[10] M. R. Schroeder and B. Atal, "Computer Simulation of Sound Transmission in Rooms," in *IEEE Int. Cony. Rec.*, pp. 150–155, 1963.

[11] M. Kleiner, B.-I. Dalenbäck, and U. P. Svensson, "Auralization – An overview," *J. Audio Eng. Soc.*, vol. 41, pp. 861 – 875, Nov. 1993.

[12] H. Lehnert and J. Blauert, "Principles of binaural room simulation," *Applied Acoustics*, vol. 36, no. 3-4, pp. 259–291, 1992.

[13] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating Interactive Virtual Acoustic Environments," *J. Audio Eng. Soc.*, vol. 49, Sept. 1999.

[14] T. Lokki, *Physically-based auralization: design, implementation, and evaluation*. PhD thesis, Helsinki University of Technology, Espoo, Finland, 2002.

[15] J. D. Miller and E. M. Wenzel, "Recent Developments in SLAB: A Software-based system for interactive spatial sound synthesis," in *International Conference on Auditory Display*, (Kyoto, Japan), July 2002.

[16] D. Schröder and M. Vorländer, "RAVEN: A Real-Time Framework for the Auralization of Interactive Virtual Environments," in *Forum Acusticum*, (Aalborg, Denmark), pp. 1541–1546, July 2011.

[17] T. Wendt, S. van de Par, and S. Ewert, "A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation," *J. Audio Eng. Soc.*, vol. 62, pp. 748–766, Nov. 2014.

[18] J. Blauert, H. Lehnert, W. Pompetzki, and N. Xiang, "Binaural Room Simulation," *Acustica*, vol. 70, Apr. 1990.

[19] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," *J. Acoust. Soc. Am.*, vol. 138, pp. 708–730, Aug. 2015.

[20] J. Meyer, *Numerical and perceptual evaluations of finite-difference time-domain simulations for room acoustics applications*. PhD thesis, Aalto University, Espoo, Finland, 2022.

[21] F. Pind, *Wave-Based Virtual Acoustics*. PhD thesis, Technical University of Denmark, Kopenhagen, Denmark, 2020.

[22] 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.*, vol. 145, pp. 2746–2760, Apr. 2019.

[23] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," in *108th Audio Engineering Society Convention*, (Paris, France), Nov. 2000.

[24] S. Müller and P. Massarani, "Transfer-Function Measurement with Sweeps," *J Audio Eng Soc*, vol. 49, pp. 443–471, June 2001.

[25] B. Bernschütz, *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording*. PhD thesis, Technical University of Berlin, Berlin, Germany, 2016.

[26] A. Lindau, *Binaural Resynthesis of Acoustical Environments. Technology and Perceptual Evaluation*. PhD thesis, Technische Universität Berlin, Berlin, Germany, 2014.

[27] C. W. Pike, *Evaluating the Perceived Quality of Binaural Technology*. PhD Thesis, University of York, York, UK, 2019.

[28] B. Rafaely, V. Tourbabin, E. Habets, Z. Ben-Hur, H. Lee, H. Gamper, L. Arbel, L. Birnie, T. Abhayapala, and P. Samarasinghe, "Spatial audio signal processing for binaural reproduction of recorded acoustic scenes – review and challenges," *Acta Acust.*, vol. 6, Oct. 2022.

[29] A. Lindau and F. Brinkmann, "Perceptual evaluation of individual headphone compensation in binaural synthesis based on non-individual recordings," *J Audio Eng Soc*, vol. 60, pp. 54–62, Sept. 2010.

[30] B. Masiero and J. Fels, "Perceptually robust headphone equalization for binaural reproduction," in *130th Audio Engineering Society Convention*, May 2011.

[31] J. G. Bolaños, A. Mäkivirta, and V. Pulkki, "Automatic Regularization Parameter for Headphone Transfer Function Inversion," *J. Audio Eng. Soc.*, vol. 64, pp. 752–761, Oct. 2016.

[32] F. Wefers, *Partitioned convolution algorithms for real-time auralization*. PhD thesis, RWTH, Aachen, Germany, 2015.

[33] B. Rafaely, *Fundamentals of spherical array processing*. New York, NY: Springer Berlin Heidelberg, 2014.

[34] J. Ahrens, H. Helmholz, D. L. Alon, and S. V. Amengual Gari, "Spherical Harmonic Decomposition of a Sound Field Using Microphones on a Circumferential Contour Around a Non-Spherical Baffle," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 3110–3119, Sept. 2022.

[35] L. McCormack, *Parametric reproduction of microphone array recordings*. PhD thesis, Aalto University, Espoo, Finland, 2023.

[36] L. Mccormack, N. Meyer-Kahlen, D. L. Alon, Z. Ben-Hur, S. V. Amengual Garí, and P. Robinson, "Six-Degrees-of-Freedom Binaural Reproduction of Head-Worn Microphone Array Capture," *J. Audio Eng. Soc.*, vol. 71, pp. 638–649, Oct. 2023.

[37] P. Majdak, "Multiple Exponential Sweep Method for Fast Measurement of Head-Related Transfer Functions," *J. Audio Eng. Soc.*, vol. 55, July 2007.

[38] B. Xie, *Head-related transfer function and virtual auditory display*. Plantation, FL: J. Ross Publishing, 2nd ed., 2013.

[39] K. Iida, *Head-Related Transfer Function and Acoustic Virtual Reality*. Singapore: Springer Singapore, 2019.

[40] "ARI HRTF-database." https://www.oeaw.ac.at/isf/das-institut/software/hrtf-database.

[41] F. Brinkmann, M. Dinakaran, R. Pelzer, J. Wohlgemuth, F. Seipl, and S. Weinzierl, "The HUTUBS HRTF database," 2019. https://api-depositonce.tu-berlin.de/server/api/core/bitstreams/8f6e24a2-1c75-4f84-a50f-74a34bb480c7/content.

[42] I. Engel, R. Daugintis, T. Vicente, A. O. Hogg, J. Pauwels, A. J. Tournier, and L. Picinali, "The SONICOM HRTF Dataset," *J. Audio Eng. Soc.*, vol. 71, pp. 241–253, May 2023.

[43] R. Sridhar, J. G. Tylka, and E. Y. Choueiri, "A database of head-related transfer function and morphological measurements," in *143th Audio Engineering Society Convention*, (New York, NY, USA), Oct. 2017.

[44] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database," *Appl. Sci.*, vol. 8, p. 2029, Oct. 2018.

[45] R. Bomhardt, M. De La Fuente Klein, and J. Fels, "A high-resolution head-related transfer function and three-dimensional ear model database," in *International Congress on Acoustics*, (Buenos Aires, Argentina), Dec. 2016.

[46] P. Majdak, F. Zotter, F. Brinkmann, J. De Muynke, M. Mihocic, and M. Noisternig, "Spatially oriented format for acoustics 2.1: Introduction and recent advances," *J. Audio Eng. Soc.*, vol. 70, pp. 565–584, July 2022.

[47] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, pp. 111–123, July 1993.

[48] C. Mendonça, "A review on auditory space adaptations to altered head-related cues," *Front. Neurosci.*, vol. 8, July 2014.

[49] P. Lladó, K. Pollack, and N. Meyer-Kahlen, "Towards a standard listener-independent HRTF to facilitate long-term adaptation," *J. Audio Eng. Soc.*, vol. 72, pp. 188 – 192, Apr. 2023.

[50] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, pp. 456–466, June 1997.

[51] B. F. G. Katz and M. Noisternig, "A comparative study of interaural time delay estimation methods," *J. Acoust. Soc. Am.*, vol. 135, pp. 3530–3540, June 2014.

[52] J.-M. Jot and V. Larcher, "Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony," in *98th Audio Engineering Society Convention*, (Paris, France), 1995.

[53] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*. Cambridge, Massachusetts: AP Professional, 1st ed., 1994.

[54] J. M. Arend, C. Pörschmann, S. Weinzierl, and F. Brinkmann, "Magnitude-Corrected and Time-Aligned Interpolation of Head-Related Transfer Functions," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3783–3799, Sept. 2023.

[55] C. Salvador, S. Sakamoto, J. Treviño, and Y. Suzuki, "Design theory for binaural synthesis: Combining microphone array recordings and head-related transfer function datasets," *Acoust. Sci. & Tech.*, vol. 38, no. 2, pp. 51–62, 2017.

[56] L. Madmoni, J. Donley, V. Tourbabin, and B. Rafaely, "Beamforming-based Binaural Reproduction by Matching of Binaural Signals," in *AES Conference on Audio for Virtual and Augmented Reality*, (Redmond, WA, USA), Aug. 2020.

[57] E. Rasumow, M. Hansen, S. van de Par, D. Puschel, V. Mellert, S. Doclo, and M. Blau, "Regularization Approaches for Synthesizing HRTF Directivity Patterns," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, pp. 215–225, Feb. 2016.

[58] L. Madmoni, J. Donley, V. Tourbabin, and B. Rafaely, "Binaural Reproduction From Microphone Array Signals Incorporating Head-Tracking," in *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, (Bologna, Italy), Sept. 2021.

[59] M. A. Gerzon, "Periphony: With-Height Sound Reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.

[60] J. Daniel, *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. PhD thesis, l'Université Paris, Paris, France, 2001.

[61] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, vol. 19 of *Springer Topics in Signal Processing*. Springer International Publishing, 2019.

[62] C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi, "ambiX - A suggested Ambisonics Format," in *Ambisonics Symposium*, July 2011.

[63] A. Politis and H. Gamper, "Comparing modeled and measurement-based spherical harmonic encoding filters for spherical microphone arrays," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, NY, USA), pp. 224–228, Oct. 2017.

[64] J. Ivanic and K. Ruedenberg, "Rotation Matrices for Real Spherical Harmonics. Direct Determination by Recursion," *J. Phys. Chem.*, vol. 100, no. 15, pp. 6342–6347, 1996.

[65] T. McKenzie, D. Murphy, and G. Kearney, "Diffuse-Field Equalisation of Binaural Ambisonic Rendering," *Appl. Sci.*, vol. 8, Oct. 2018.

[66] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely, "Efficient Representation and Sparse Sampling of Head-Related Transfer Functions Using Phase-Correction Based on Ear Alignment," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, pp. 2249–2262, Dec. 2019.

[67] 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.*, vol. 143, pp. 3616–3627, June 2018.

[68] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *DAGA - Fortschritte der Akustik*, (Munich, Germany), Mar. 2018.

[69] C. Hold, N. Meyer-Kahlen, and V. Pulkki, "Magnitude-Least-Squares Binaural Ambisonic Rendering with Phase Continuation," in *DAGA - Fortschritte der Akustik*, (Hamburg, Germany), Mar. 2023.

[70] "IEM Plugin Suite." https://git.iem.at/audioplugins/IEMPluginSuite.

[71] H. Helmholz, C. Andersson, and J. Ahrens, "Real-Time Implementation of Binaural Rendering of High-Order Spherical Microphone Array Signals," in *DAGA - Fortschritte der Akustik*, (Rostock, Germany), Mar. 2019.

[72] H. Helmholz and T. Lübeck, "Updates on the Real-Time Spherical Array Renderer (ReTiSAR)," in *DAGA - Fortschritte der Akustik*, 2020.

[73] N. Meyer-Kahlen, P. Piiroinen, G. Vishwanath, P. Juntunen, E. Tiainen, and S. J. Schlecht, "Inside The Quartet - A first-person virtual reality string quartet production," (Espoo, Finland), May 2023.

[74] T. Deppisch, N. Meyer-Kahlen, B. Hofer, T. Łatka, and T. Żernicki, "HOAST: A Higher-Order Ambisonics Streaming Platform," in *148Ath Audio Engineering Society Convention*, 2020.

[75] J. Pätynen, S. Tervo, and T. Lokki, "Analysis of concert hall acoustics via visualizations of time-frequency and spatiotemporal responses," *J. Acoust. Soc. Am.*, vol. 133, pp. 842–857, Feb. 2013.

[76] W. Lachenmayr, N. Meyer-Kahlen, O. Colella Gomes, A. Kuusinen, and T. Lokki, "Chamber music hall acoustics: Measurements and perceptual differences," *J. Acoust. Soc. Am.*, vol. 154, pp. 388–400, July 2023.

[77] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial Decomposition Method for Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 61, pp. 17–28, Mar. 2013.

[78] 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.*, vol. 68, pp. 368–354, May 2020.

[79] A. Pawlak, H. Lee, T. Lund, and A. Mäkivirta, "Subjective evaluation of spatial analysis and synthesis methods using different microphone arrays," in *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, (Bologna, Italy), Sept. 2021.

[80] A. Pawlak, H. Lee, A. Mäkivirta, and T. Lund, "Spatial Analysis and Synthesis Methods: Subjective and Objective Evaluations Using Various Microphone Arrays in the Auralization of a Critical Listening Room," Jan. 2024. arXiv:2401.15023 [cs, eess].

[81] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: analysis and synthesis," *J. Audio Eng. Soc.*, vol. 53, pp. 1115–1127, Dec. 2005.

[82] M. Zaunschirm, M. Frank, and F. Zotter, "Binaural Rendering with measured room responses: first-order ambisonic microphone vs. dummy head," *Appl. Sci.*, vol. 10, Feb. 2020.

[83] P. Stade, J. Arend, and C. Pörschmann, "A parametric model for the synthesis of binaural room impulse responses," in *173rd Meeting of Acoustical Society of America and 8th Forum Acusticum*, (Boston, Massachusetts), Aug. 2017.

[84] P. Coleman, P. J. B. Jackson, L. Remaggi, and F. Melchior, "Object-Based Reverberation for Spatial Audio," *J. Audio Eng. Soc.*, vol. 65, pp. 66–77, Feb. 2017.

[85] L. McCormack, N. Meyer-Kahlen, and A. Politis, "Multi-directional parameterisation and rendering of spatial room impulse responses," in *24th International Congress on Acoustics*, (Gyeongju, Korea), Oct. 2022.

[86] J. Riionheimo, "Movie Sound, Part 2: Preference and Attribute Ratings of Six Listening Environments," *J. Audio Eng. Soc.*, vol. 69, p. 12, Feb. 2021.

[87] J. Pätynen, S. Tervo, and T. Lokki, "Amplitude panning decreases spectral brightness with concert hall auralizations," in *AES 55th International Conference*, Aug. 2014.

[88] S. Tervo, J. Pätynen, N. Kaplanis, M. Lydolf, S. Bech, and T. Lokki, "Spatial analysis and synthesis of car audio system and car cabin acoustics with a compact microphone array," *J. Audio Eng. Soc.*, vol. 63, pp. 914–925, Feb. 2015.

[89] Y. Yamasaki and T. Itow, "Measurement of spatial information in sound fields by closely located four point microphone method.," *J. Acoust. Soc. Jpn (E)*, vol. 10, no. 2, pp. 101–110, 1989.

[90] N. Meyer-Kahlen, S. V. Amengual Garí, and T. Lokki, "What the spatial decomposition method can and cannot do," in *24th International Congress on Acoustics*, (Gyeongju, Korea), Oct. 2022.

[91] M. Frank and F. Zotter, "Spatial impression and directional resolution in the reproduction of reverberation," in *DAGA - Fortschritte der Akustik*, (Aachen, Germany), Mar. 2016.

[92] M. Zaunschirm, M. Frank, and F. Zotter, "BRIR synthesis using first-order microphone arrays," in *144th Audio Engineering Society Convention*, May 2018.

[93] M. Zaunschirm, F. Zagala, and F. Zotter, "Auralization of High-Order Directional Sources from First-Order RIR Measurements," *Appl. Sci.*, vol. 10, May 2020.

[94] K. Prawda, S. J. Schlecht, and V. Välimäki, "Evaluation of Reverberation Time Models with Variable Acoustics," in *Proceedings of the 17th Sound and Music Computing Conference*, June 2020.

[95] H. Kuttruff, *Room acoustics*. London, England; New York, NY: Spon Press, 4th ed., 2000.

[96] B. Alary, P. Massé, V. Välimäki, and M. Noisternig, "Assessing the anisotropic features of spatial impulse responses," in *EAA Spatial Audio Signal Processing Symposium,*, pp. 43–48, Sept. 2019.

[97] M. Berzborn and M. Vorländer, "Directional sound field decay analysis in performance spaces," *Building Acoustics*, vol. 28, pp. 249–263, Sept. 2021.

[98] M. Nolan, M. Berzborn, and E. Fernandez-Grande, "Isotropy in decaying reverberant sound fields," *J. Acoust. Soc. Am.*, vol. 148, pp. 1077–1088, Aug. 2020.

[99] J. Daniel and S. Kitić, "Time Domain Velocity Vector for Retracing the Multipath Propagation," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Barcelona, Spain), pp. 421–425, May 2020.

[100] N. Meyer-Kahlen, S. J. Schlecht, and T. Lokki, "Parametric late reverberation from broadband directional estimates," in *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, (Bologna, Italy), Sept. 2021.

[101] S. V. Amengual Garí, P. T. Calamia, and P. W. Robinson, "Optimizations of the spatial decomposition method for binaural reproduction," *J. Audio Eng. Soc.*, vol. 68, pp. 959–976, Dec. 2020.

[102] E. Hoffbauer and M. Frank, "4-Directional Ambisonic Spatial Decomposition Method with Reduced Temporal Artifacts," *J. Audio Eng. Soc.*, vol. 70, pp. 1002–1014, Dec. 2022.

[103] U. Sloma, N. Merten, T. Thron, K. Brandenburg, F. Wollwert, R. Profeta, and C. Rodriguez, "Proof of concept of a binaural renderer with increased plausibility," in *DAGA - Fortschritte der Akustik*, (Hamburg, Germany), Mar. 2023.

[104] J. Ahrens, "Perceptual Evaluation of Binaural Auralization of Data Obtained from the Spatial Decomposition Method," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, NY, USA), Oct. 2019.

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

[106] A. Lindau, "The Perception of System Latency in Dynamic Binaural Synthesis," in *NAG/DAGA - Fortschritte der Akustik*, (Rotterdam, NL), Jan. 2009.

[107] N. Meyer-Kahlen, S. V. Amengual Garí, T. McKenzie, S. J. Schlecht, and T. Lokki, "Transfer-plausibility of binaural rendering with different real-world references," in *DAGA - Fortschritte der Akustik*, Mar. 2022.

[108] F. Brinkmann, R. Roden, A. Lindau, and S. Weinzierl, "Audibility and Interpolation of Head-Above-Torso Orientation in Binaural Technology," *IEEE J. Sel. Top. Signal Process.*, vol. 9, pp. 931–942, Aug. 2015.

[109] P.-Q. David, P. Stitt, and B. Katz, "RoomZ: Spatial panning plugin for dynamic auralisations based on RIR convolution," in *AES International Conference on Spatial and Immersive Audio*, (Huddersfield, UK), Aug. 2023.

[110] G. Götz, S. J. Schlecht, A. M. Ornelas, and V. Pulkki, "Autonomous Robot Twin System for Room Acoustic Measurements," *J. Audio Eng. Soc.*, vol. 69, pp. 261–272, Apr. 2021.

[111] G. Götz, I. Ananthabhotla, S. Amengual Garí, and P. Calamia, "Autonomous Room Acoustic Measurements using Rapidly-Exploring Random Trees and Gaussian Processes," in *Proceedings of the 10th Convention of the European Acoustics Association: Forum Acusticum*, (Torino, Italy), pp. 1655–1662, Jan. 2024.

[112] G. Stolz, F. Klein, S. Werner, L. Treybig, A. Bley, and C. Martin, "Discussion of Acoustic and Perceptual Optimization Methods for Measuring Spatial Room Impulse Responses with a Mobile Robotic Platform," in *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, (Bologna, Italy), Sept. 2023.

[113] S. Werner, F. Klein, A. Neidhardt, U. Sloma, C. Schneiderwind, and K. Brandenburg, "Creation of Auditory Augmented Reality Using a Position-Dynamic Binaural Synthesis System—Technical Components, Psychoacoustic Needs, and Perceptual Evaluation," *Appl. Sci.*, vol. 11, Jan. 2021.

[114] O. Puomio, N. Meyer-Kahlen, and T. Lokki, "Locating Image Sources from Multiple Spatial Room Impulse Responses," *Appl. Sci.*, vol. 11, Mar. 2021.

[115] A. Geldert, N. Meyer-Kahlen, and S. J. Schlecht, "Interpolation of Spatial Room Impulse Responses Using Partial Optimal Transport," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Rhodes Island, Greece), June 2023.

[116] K. Müller and F. Zotter, "Auralization based on multi-perspective ambisonic room impulse responses," *Acta Acust.*, vol. 4, Nov. 2020.

[117] T. McKenzie, N. Meyer-Kahlen, C. Hold, S. J. Schlecht, and V. Pulkki, "Auralization of Measured Room Transitions in Virtual Reality," *J. Audio Eng. Soc.*, vol. 71, pp. 326–337, June 2023.

[118] K. Brandenburg, E. Cano, F. Klein, T. Köllmer, H. Lukashevich, U. Sloma, and S. Werner, "Plausible Augmentation of Auditory Scenes Using Dynamic Binaural Synthesis for Personalized Auditory Realities," in *AES Conference on Audio for Virtual and Augmented Reality*, (Redmond, WA, USA), Aug. 2018.

[119] H. Gamper and T. Lokki, "Audio Augmented Reality in Telecommunication through virtual auditory display," in *16th International Conference on Auditory Display (ICAD-2010)*, June 2010.

[120] R. Albrecht, *Methods and applications of mobile audio augmented reality*. PhD thesis, Aalto University, Espoo, Finland, 2016.

[121] T. Lokki, H. Nironen, S. Vesa, L. Savioja, A. Härmä, and M. Karjalainen, "Application Scenarios of Wearable and Mobile Augmented Reality Audio," in *116th Audio Engineering Society Convention*, (Berlin, Germany), May 2004.

[122] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented Reality Audio for Mobile and Wearable Appliances," *J. Audio Eng. Soc.*, vol. 52, p. 23, June 2004.

[123] J. Rämö and V. Välimäki, "Digital Augmented Reality Audio Headset," *Journal of Electrical and Computer Engineering*, vol. 2012, Oct. 2012.

[124] R. Gupta, J. He, R. Ranjan, W.-S. Gan, F. Klein, C. Schneiderwind, A. Neidhardt, K. Brandenburg, and V. Valimaki, "Augmented/Mixed Reality Audio for Hearables: Sensing, control, and rendering," *IEEE Signal Process. Mag.*, vol. 39, pp. 63–89, May 2022.

[125] H. E.-B. Zidan, U. P. Svensson, and J. L. Nielsen, "Room acoustical parameters of two electronically connected rooms," *J. Acoust. Soc. Am.*, vol. 138, pp. 2235–2245, Oct. 2015.

[126] A. Haeussler and S. van de Par, "Crispness, speech intelligibility, and coloration of reverberant recordings played back in another reverberant room (Room-In-Room)," *J. Acoust. Soc. Am.*, vol. 145, pp. 931–944, Feb. 2019.

[127] P. A. Naylor and N. D. Gaubitch, eds., *Speech Dereverberation*. Signals and Communication Technology, London: Springer, 2010.

[128] T. Deppisch, N. Meyer-Kahlen, and S. V. A. Garí, "Blind Identification of Binaural Room Impulse Responses from Smart Glasses," Mar. 2024. arXiv:2403.19217 [eess].

[129] "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969. APPENDIX C 1965 Revised List of Phonetically Balanced Sentences (Harvard Sentences).

[130] A. Hofmann, N. Meyer-Kahlen, and S. J. Schlecht, "Audiovisual Congruence and Localization Performance in Virtual Reality: 3D Loudspeaker Model vs. Human Avatar," *J. Audio Eng. Soc.*, 2024. accepted.

[131] 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*, vol. 26, Jan. 2022.

[132] S. V. Amengual Garí, P. W. Robinson, and P. T. Calamia, "Room acoustic characterization for binaural rendering: From spatial room impulse responses to deep learning," in *International Congress on Acoustics*, (Gyeongju, Korea), Oct. 2022.

[133] H. Kon and H. Koike, "An auditory scaling method for reverb synthesis from a single two-dimensional image," *Acoust. Sci. & Tech.*, vol. 41, pp. 675–685, July 2020.

[134] N. Singh, J. Mentch, J. Ng, M. Beveridge, and I. Drori, "Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis," Aug. 2021. arXiv:2103.14201 [cs, eess].

[135] H. Kim, L. Remaggi, P. J. Jackson, and A. Hilton, "Immersive Spatial Audio Reproduction for VR/AR Using Room Acoustic Modelling from 360° Images," in *Conference on Virtual Reality and 3D User Interfaces (VR)*, (Osaka, Japan), pp. 120–126, Mar. 2019.

[136] L. Remaggi, H. Kim, A. Neidhardt, A. Hilton, and J. B. Jackson, "Perceived quality and spatial impression of room reverberation in VR reproduction from measured images and acoustics," in *International Congress on Acoustics*, (Aachen, Germany), Sept. 2019.

[137] C. J. Steinmetz, V. K. Ithapu, and P. T. Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, NY, USA), Oct. 2021.

[138] A. Ratnarajah, I. Ananthabhotla, V. K. Ithapu, P. Hoffmann, D. Manocha, and P. Calamia, "Towards Improved Room Impulse Response Estimation for Speech Recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Rhodes Island, Greece), June 2023.

[139] 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," in *Interspeech*, (Dublin, Ireland), Aug. 2023.

[140] S. Lee, H.-S. Choi, and K. Lee, "Yet Another Generative Model for Room Impulse Response Estimation," in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, NY, USA), Oct. 2023.

[141] S. Lee, H.-S. Choi, and K. Lee, "Differentiable Artificial Reverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, 2022.

[142] J. Su, Z. Jin, and A. Finkelstein, "Acoustic Matching By Embedding Impulse Responses," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Barcelona, Spain), May 2020.

[143] K. Lee, J. Seo, K. Choi, S. Lee, and B. S. Chon, "Room Impulse Response Estimation in a Multiple Source Environment," in *AES International Conference on Spatial and Immersive Audio*, (Huddersfield, UK), Aug. 2023.

[144] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Acoustic Characterization of Environments (ACE) Challenge Results Technical Report," June 2017. arXiv:1606.03365 [cs].

[145] H. Gamper and I. J. Tashev, "Blind Reverberation Time Estimation Using a Convolutional Neural Network," in *16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, (Tokyo, Japan), pp. 136–140, Sept. 2018.

[146] 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.*, vol. 153, pp. 3532–3542, June 2023.

[147] G. Xu, L. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Trans. Signal Process.*, vol. 43, pp. 2982–2993, Dec. 1995.

[148] T. Mei, P. Hang, and A. Mertins, "Adaptive estimation and reshaping of room impulse response," *Int. J. Speech Technol.*, vol. 18, pp. 91–95, Mar. 2015.

[149] B. Jo and P. T. Calamia, "Robust blind multichannel identification based on a phase constraint and different lp-norm constraints," in *Eur. Signal Process. Conf (EUSIPCO)*, (Amsterdam, Netherlands), pp. 1966–1970, Jan. 2021.

[150] M. Schneider and W. Kellermann, "The generalized frequency-domain adaptive filtering algorithm as an approximation of the block recursive least-squares algorithm," *EURASIP J. Adv. Signal Process.*, vol. 2016, Dec. 2016.

[151] T. Deppisch, J. Ahrens, S. Amengual Garí, and P. Calamia, "Blind Estimation of Spatial Room Impulse Responses Using a Pseudo Reference Signal," in *Hands-free Speech Communication and Microphone Arrays at ICASSP*, (Seoul, South Korea), Apr. 2024. accepted.

[152] Y. Hu, S. Gannot, and T. D. Abhayapala, "Generalized Relative Harmonic Coefficients," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Rhodes Island, Greece), June 2023.

[153] S. Kitić and J. Daniel, "Blind identification of Ambisonic reduced room impulse response," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 32, pp. 443–458, 2024.

[154] J. M. Arend, S. V. Amengual Garí, C. Schissler, F. Klein, and P. W. Robinson, "Six-degrees-of-freedom parametric spatial audio based on one monaural room impulse response," *J. Audio Eng. Soc.*, vol. 69, pp. 557–575, Nov. 2021.

[155] A. Neidhardt, A. I. Tommy, and A. D. Pereppadan, "Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets," in *144th Audio Engineering Society Convention*, (Milano, Italy), May 2018.

[156] A. Neidhardt and A. M. Zerlik, "The Availability of a Hidden Real Reference Affects the Plausibility of Position-Dynamic Auditory AR," *Front. Virtual Real.*, vol. 2, Sept. 2021.

[157] C. Pike, F. Melchior, and T. Tew, "Assessing the Plausibility of Non-Individualised Dynamic Binaural Synthesis in a Small Room," in *AES 55th International Conference*, (Helsinki, Finland), Aug. 2014.

[158] A. Lindau and S. Weinzierl, "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acust.*, vol. 98, pp. 804–810, Sept. 2012.

[159] N. Meyer-Kahlen, D. Rudrich, S. A. Wirler, S. Winther, and M. Frank, "DIY Modifications for Transparent Headphones," in *148th Audio Engineering Society Convention*, 2020.

[160] N. Klanjscek, L. David, and M. Frank, "Evaluation of an E-Learning Tool for Augmented Acoustics in Music Education," *Music & Science*, vol. 4, Jan. 2021.

[161] F. Schultz, A. Lindau, M. Makarski, and S. Weinzierl, "An extraaural headphone for optimized binaural reproduction," in *26th Tonmeistertagung*, pp. 702–714, Nov. 2010.

[162] A. Mülleder and F. Zotter, "Ultralight circumaural open headphones," in *154th Audio Engineering Society Convention*, (Espoo, Finland), May 2023.

[163] A. Mülleder, M. Romanov, N. Meyer-Kahlen, and F. Zotter, "Do-it-yourself headphones and development platform for augmented-reality audio," in *AES Conference on Immersive and Spatial Audio*, (Huddersfield, UK), Aug. 2023.

[164] T. 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.*, vol. 63, pp. 799 – 810, Oct. 2015.

[165] R. Baumgartner, P. Majdak, and B. Laback, "Modeling Sound-Source Localization in Sagittal Planes for Human Listeners," *J. Acoust. Soc. Am.*, vol. 136, pp. 791–802, Aug. 2014.

[166] P. Majdak, C. Hollomey, and R. Baumgartner, "AMT 1.x: A toolbox for reproducible research in auditory modeling," *Acta Acust.*, vol. 6, May 2022.

[167] "ITU-R BS.1116-3: Methods for the subjective assessment of small impairments in audio sys- tems," 2015.

[168] "ITU-R BS.1534-3: Methods for the subjective assessment of intermediate quality level of audio systems," 2015.

[169] J. Herre and S. Disch, "MPEG-I Immersive Audio – Reference Model For The Virtual/Augmented Reality Audio Standard," *J. Audio Eng. Soc.*, vol. 71, pp. 229–240, May 2023.

[170] S. Weinzierl and F. Brinkman, "Audio Quality Assessment for Virtual Reality," in *Sonic Interactions in Virtual Environments* (M. Geronazzo and S. Serafin, eds.), Human–Computer Interaction Series, pp. 145–178, Cham, Switzerland: Springer International Publishing, 2023.

[171] W. Pompetzki and J. Blauert, "A study on the perceptual authenticity of binaural room sumulation," in *Wallace Clement Sabine Centennial Symposium*, (Cambridge, MA, USA), June 1994.

[172] M. Aretz, *Combined wave and ray based room acoustic simulations of small rooms*. PhD thesis, RWTH, Aachen, Germany, 2012.

[173] B. F. G. Katz, D. Poirier-Quinot, B. N. J. Postma, D. Thery, U. Paris-Saclay, and P. Luizard, "Objective and perceptive evaluations of high-resolution room acoustic simulations and auralizations," in *Euronoise*, (Crete, Greece), 2018.

[174] S. Fichna, C. Kirsch, B. U. Seeber, and S. D. Ewert, "Perceptual evaluation of simulated and real acoustic scenes with different acoustic level of detail," in *24th International Congress on Acoustics*, Oct. 2022.

[175] V. Pulkki and J. Merimaa, "Spatial impulse response rendering II: reproduction of diffuse sound and listening tests," *J. Audio Eng. Soc.*, vol. 54, pp. 3–20, Jan. 2006.

[176] P. Zahorik, F. Wightman, and D. Kistler, "On the discriminability of virtual and real sound sources," in *Workshop on Applications of Signal Processing to Audio and Accoustics*, (New Paltz, NY, USA), pp. 76–79, 1995.

[177] W. M. Hartmann and A. Wittenberg, "On the externalization of sound images," *J. Acoust. Soc. Am.*, vol. 99, pp. 3678–3688, June 1996.

[178] E. H. A. Langendijk and A. W. Bronkhorst, "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *J. Acoust. Soc. Am.*, vol. 107, pp. 528–537, Jan. 2000.

[179] E. Thompson, "Machines, Music, and the Quest for Fidelity: Marketing the Edison Phonograph in America, 1877–1925," *Musical Quarterly*, vol. 79, no. 1, pp. 131–171, 1995.

[180] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A Spatial Audio Quality Inventory (SAQI)," *Acta Acustica united with Acustica*, vol. 100, pp. 984–994, Sept. 2014.

[181] N. Zacharov, T. Pedersen, and C. Pike, "A common lexicon for spatial sound quality assessment - latest developments," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, (Lisbon, Portugal), June 2016.

[182] Y. Wycisk, K. Sander, R. Kopiez, F. Platz, S. Preihs, and J. Peissig, "Wrapped into sound: Development of the Immersive Music Experience Inventory (IMEI)," *Front. Psychol.*, vol. 13, Sept. 2022.

[183] M. Blau, A. Budnik, M. Fallahi, H. Steffens, S. D. Ewert, and S. van de Par, "Toward realistic binaural auralizations – perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario," *Acta Acust.*, vol. 5, Jan. 2021.

[184] F. Stärz, L. O. H. Kroczek, S. Roßkopf, A. Mühlberger, S. van de Par, and M. Blau, "Comparing Room Acoustical Ratings In An Interactive Virtual Environment To Those In The Real Room," in *10th Convention of the European Acoustics Association: Forum Acusticum*, (Torino, Italy), Sept. 2023.

[185] K. Enge, M. Frank, and R. Höldrich, "Listening experiment on the plausibility of acoustic modeling in virtual reality," in *DAGA - Fortschritte der Akustik*, 2020.

[186] C. Schneiderwind, M. Richter, A. Neidhardt, and N. Merten, "Effects of Modified Late Reverberation on Audio-Visual Plausibility and Externalization in AR," in *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, (Bologna, Italy), 2023.

[187] M. Gospodarek, O. Warusfel, P. Ripollés, and A. Roginska, "Methodology for perceptual evaluation of plausibility with self-translation of the listener," in *AES International Conference on Audio for Virtual and Augmented Reality*, (Redmond, WA, USA), Aug. 2022.

[188] C. Kuhn-Rahloff, *Realitätstreue, Natürlichkeit, Plausibilität*. Springer Heidelberg, 2012.

[189] V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopčo, "Sound Externalization: A Review of Recent Research," *Trends in Hearing*, vol. 24, Jan. 2020.

[190] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, (Lisbon, Portugal), June 2016.

[191] J. Oberem, B. Masiero, and J. Fels, "Experiments on authenticity and plausibility of binaural reproduction via headphones employing different recording methods," *Applied Acoustics*, vol. 114, pp. 71–78, Dec. 2016.

[192] A. Genovese, "Individualisation and reverberation factors in the subjective assessment of plausibility in a binaural auditory display," 2014. University of York, York, UK, Master's Thesis.

[193] J. Blauert, *Spatial Hearing*. MIT Press, Cambridge, Massachusetts, 1983.

[194] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proc. Natl. Acad. Sci. USA*, vol. 113, pp. E7856–E7865, Nov. 2016.

[195] T. McKenzie, N. Meyer-Kahlen, and S. J. Schlecht, "The role of source signal similarity in distinguishing between different positions in a room," in *AES Conference on Immersive and Spatial Audio*, (Huddersfield, UK), Aug. 2023.

[196] A. Kuusinen and T. Lokki, "Recognizing individual concert halls is difficult when listening to the acoustics with different musical passages," *J. Acoust. Soc. Am.*, vol. 148, pp. 1380–1390, Sept. 2020.

[197] J. T. Wixted, E. Vul, L. Mickes, and B. M. Wilson, "Models of lineup memory," *Cognitive Psychology*, vol. 105, pp. 81–114, Sept. 2018.

[198] G. C. Stecker, T. M. Moore, M. Folkerts, D. Zotkin, and R. Duraiswami, "Toward objective measures of auditory co-immersion in virtual and augmented reality," in *AES Conference on Audio for Virtual and Augmented Reality*, (Redmond, WA, USA), Aug. 2018.

[199] 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 Trans. Visual. Comput. Graphics*, vol. 29, pp. 4472–4481, Nov. 2023.

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

[201] M. Frank and D. Perinovic, "Matching auditory and visual room size, distance, and source orientation in virtual reality," in *AudioMostly 2022*, (St. Pölten Austria), pp. 80–83, Sept. 2022.

[202] B. Burnett, A. Neidhardt, Z. Cvetković, H. Hacıhabiboğlu, and E. De Sena, "User Expectation of Room Acoustic Parameters in Virtual Reality Environments," in *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, (Bologna, Italy), Sept. 2023.

[203] R. G. Crowder, "A common basis for auditory sensory storage in perception and immediate memory," *Perception & Psychophysics*, vol. 31, pp. 477–483, Sept. 1982.

[204] M. H. Ashcraft and G. A. Radvansky, *Cognition*. Boston: Pearson Education, 6th ed., 2014.

[205] M. A. Nees, "Have We Forgotten Auditory Sensory Memory? Retention Intervals in Studies of Nonverbal Auditory Working Memory," *Front. Psychol.*, vol. 7, Dec. 2016.

[206] A. Baddeley, "Working Memory: Theories, Models, and Controversies," *Annu. Rev. Psychol.*, vol. 63, Jan. 2012.

[207] M. Nastasa, N. Meyer-Kahlen, and S. J. Schlecht, "Assessing Room Acoustic Memory using a Yes/No and a 2-AFC Paradigm," in *Nordic SMC*, Nov. 2021.

[208] F. Klein, T. Surdu, L. Treybig, and S. Werner, "The Ability to Memorize Acoustic Features in a Discrimination Task," *J. Audio Eng. Soc.*, vol. 71, pp. 254–266, May 2023.

[209] J. Heldmann and S. J. Schlecht, "The Role of Modal Excitation in Colorless Reverberation," in *24th International Conference on Digital Audio Effects (DAFx)*, (Vienna, Austria), pp. 206–213, Sept. 2021.

[210] J. Oberem, J.-G. Richter, D. Setzer, J. Seibold, I. Koch, and J. Fels, "Experiments on localization accuracy with non-individual and individual HRTFs comparing static and dynamic reproduction methods," in *DAGA - Fortschritte der Akustik*, (Munich, Germany), Mar. 2020.

[211] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *Journal of the Audio Engineering Society*, vol. 49, no. 10, pp. 904–916, 2001.

[212] N. Meyer-Kahlen, M. Kastemaa, S. J. Schlecht, and T. Lokki, "Measuring Motion-to-Sound Latency in Virtual Acoustic Rendering Systems," *J. Audio Eng. Soc.*, vol. 71, pp. 390–398, June 2023.

[213] J. Sundberg and M. Nordenberg, "Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech," *J. Acoust. Soc. Am.*, vol. 120, pp. 453–457, July 2006.

[214] C. Pörschmann and J. M. Arend, "Investigating phoneme-dependencies of spherical voice directivity patterns II: Various groups of phonemes," *J. Acoust. Soc. Am.*, vol. 153, pp. 179–190, Jan. 2023.

[215] C. Pörschmann and J. M. Arend, "Phoneme dependence of horizontal asymmetries in voice directivity," *JASA Express Letters*, vol. 4, Feb. 2024.

[216] T. Lübeck and C. Pörschmann, "Evaluating the Plausibility of Non-Individual Head-Related Transfer Functions in Anechoic Conditions," in *10th Convention of the European Acoustics Association: Forum Acusticum*, (Torino, Italy), Sept. 2023.

[217] J. I. Engel Alonso-Martinez, *Improving binaural audio techniques for augmented reality*. PhD thesis, Imperial College London, London, UK, May 2021.

Aalto-DT 139/2024

ISBN 978-952-64-1912-1 (printed)
ISBN 978-952-64-1913-8 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Department of Information and Communications Engineering
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL THESES