Efficient Binaural Rendering of Virtual Acoustic Realities

– Technical and Perceptual Concepts –

vorgelegt von
M. Sc.
Johannes Mathias Arend
ORCID: 0000-0002-5403-4076

an der Fakultät I – Geistes- und Bildungswissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
– Dr. rer. nat. –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Tilman Santarius
Gutachter:	Prof. Dr. Stefan Weinzierl
Gutachter:	Prof. Dr. Christoph Pörschmann

Tag der wissenschaftlichen Aussprache: 11. Februar 2022

Berlin 2022

Efficient Binaural Rendering of Virtual Acoustic Realities – Technical and Perceptual Concepts

Johannes Mathias Arend

Doctoral Dissertation

Technical University of Berlin – Audio Communication Group
TH Köln – University of Applied Sciences – Institute of Communications Engineering

This dissertation was written at the Technical University of Berlin – Audio Communication Group under the supervision of Prof. Dr. Stefan Weinzierl. The majority of the research work was carried out at the TH Köln – University of Applied Sciences – Institute of Communications Engineering under the supervision of Prof. Dr. Christoph Pörschmann, funded by the German Federal Ministry of Education and Research (BMBF), support code BMBF-03FH014IX5-NarDasS, and by the European Funds for Regional Development (EFRE), support code EFRE-0801444-EarKAR. Part of the research work was done as a research intern at Facebook Reality Labs Research, Redmond, WA, USA.

Acknowledgments

I would like to thank Stefan Weinzierl for supervising this dissertation. I am grateful for his valuable advice with a view to the big picture, his keen eye for detail, and his constant support in my work. I am delighted to have been able to do the work under his guidance and in collaboration with his research group at TU Berlin.

I would also like to thank my mentor and supervisor, Christoph Pörschmann, who made this work possible and initiated it. His visionary thinking and countless ideas have been inspiring and motivating and all too often set me on new paths and research topics. I am grateful that he has made it possible for me to pursue all my different research interests in a great working environment over the past six years and that we have been able to shape the focus of the research group at TH Köln together during this time.

My thanks also go to Henrich R. Liesefeld, who significantly contributed to my scientific development and guided me in scientific working and academic publishing. Many thanks for all the fruitful discussions, whether about the details of a study or the world of science in general.

I would also like to thank my former and current colleagues at TH Köln for the friendly atmosphere and the professional and inspiring teamwork (in alphabetic order): David Bau, Benjamin Bernschütz, Damian Dziwis, Aaron Finkenthei, Raphaël Gillioz, Antje Goldenberg, Narea Jantzen, Tim Lübeck, Melissa Ramírez, Philipp Stade, Arnau Vázquez-Giner. A special thanks goes to Benjamin Bernschütz, who showed me during my master studies that there is still so much to discover and explore in the audio field and thus motivated me to start a PhD.

I am also grateful for the excellent collaboration, fruitful discussions, scientific exchanges, and many nice conferences and meetings with my colleagues in virtual acoustics (in alphabetical order): David Ackermann, Jens Ahrens, David Lou Alon, Lukas Aspöck, Owen Brimijoin, Fabian Brinkmann, Paul T. Calamia, Isaac Engel, Matthias Frank, Sebastià V. Amengual Garí, Hannes Helmholz, Florian Klein, Alexander Lindau, Annika Neidhardt, Chris Pike, Henri Pöntynen, Martin Pollow, Philip W. Robinson, Christian Sander, Carl Schissler, Frank Schultz, Jonas Stienen, Stephan Werner, Franz Zotter.

Many thanks also to all who have participated in the listening experiments. It is great to see that over the years, we have been able to establish a group of expert listeners who have consistently participated in new studies.

Many dear thanks to my family for their unconditional support in all circumstances and their great understanding. I would especially like to thank Melissa Ramírez for her endless support and constant encouragement.

Abstract

Binaural rendering aims to immerse the listener in a virtual acoustic scene, making it an essential method for spatial audio reproduction in virtual or augmented reality (VR/AR) applications. The growing interest and research in VR/AR solutions yielded many different methods for the binaural rendering of virtual acoustic realities, yet all of them share the fundamental idea that the auditory experience of any sound field can be reproduced by reconstructing its sound pressure at the listener's eardrums. This thesis addresses various state-of-the-art methods for 3 or 6 degrees of freedom (DoF) binaural rendering, technical approaches applied in the context of headphone-based virtual acoustic realities, and recent technical and psychoacoustic research questions in the field of binaural technology. The publications collected in this dissertation focus on technical or perceptual concepts and methods for efficient binaural rendering, which has become increasingly important in research and development due to the rising popularity of mobile consumer VR/AR devices and applications. The thesis is organized into five research topics: Head-Related Transfer Function Processing and Interpolation, Parametric Spatial Audio, Auditory Distance Perception of Nearby Sound Sources, Binaural Rendering of Spherical Microphone Array Data, and Voice Directivity. The results of the studies included in this dissertation extend the current state of research in the respective research topic, answer specific psychoacoustic research questions and thereby yield a better understanding of basic spatial hearing processes, and provide concepts, methods, and design parameters for the future implementation of technically and perceptually efficient binaural rendering.

Zusammenfassung

Binaurales Rendering zielt darauf ab, dass der Hörer in eine virtuelle akustische Szene eintaucht, und ist somit eine wesentliche Methode für die räumliche Audiowiedergabe in Anwendungen der virtuellen Realität (VR) oder der erweiterten Realität (AR – aus dem Englischen Augmented Reality). Das wachsende Interesse und die zunehmende Forschung an VR/AR-Lösungen führte zu vielen verschiedenen Methoden für das binaurale Rendering virtueller akustischer Realitäten, die jedoch alle die grundlegende Idee teilen, dass das Hörerlebnis eines beliebigen Schallfeldes durch die Rekonstruktion seines Schalldrucks am Trommelfell des Hörers reproduziert werden kann. Diese Arbeit befasst sich mit verschiedenen modernsten Methoden zur binauralen Wiedergabe mit 3 oder 6 Freiheitsgraden (DoF – aus dem Englischen Degree of Freedom), mit technischen Ansätzen, die im Kontext kopfhörerbasierter virtueller akustischer Realitäten angewandt werden, und mit aktuellen technischen und psychoakustischen Forschungsfragen auf dem Gebiet der Binauraltechnik. Die in dieser Dissertation gesammelten Publikationen befassen sich mit technischen oder wahrnehmungsbezogenen Konzepten und Methoden für effizientes binaurales Rendering, was in der Forschung und Entwicklung aufgrund der zunehmenden Beliebtheit von mobilen Verbraucher-VR/AR-Geräten und -Anwendungen zunehmend an Relevanz gewonnen hat. Die Arbeit ist in fünf Forschungsthemen gegliedert: Verarbeitung und Interpolation von Außenohrübertragungsfunktionen, parametrisches räumliches Audio, auditive Entfernungswahrnehmung ohrnaher Schallquellen, binaurales Rendering von sphärischen Mikrofonarraydaten und Richtcharakteristik der Stimme. Die Ergebnisse der in dieser Dissertation enthaltenen Studien erweitern den aktuellen Forschungsstand im jeweiligen Forschungsfeld, beantworten spezifische psychoakustische Forschungsfragen und führen damit zu einem besseren Verständnis grundlegender räumlicher Hörprozesse, und liefern Konzepte, Methoden und Gestaltungsparameter für die zukünftige Umsetzung eines technisch und wahrnehmungsbezogen effizienten binauralen Renderings.

List of included publications

This thesis is a cumulative dissertation. For the sake of conciseness, the print version includes only the key publications (co-)written by the present author on the various research topics he worked on as part of his PhD. The complete list of publications for this cumulative dissertation, which is appended at the end of this thesis, contains further publications (co-)written by the author during his PhD period.

Head-Related Transfer Function Processing and Interpolation

Arend, J. M., Brinkmann, F., & Pörschmann, C. (2021). Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions. *J. Audio Eng. Soc.*, *69*(1/2), 104–117. https://doi.org/10.17743/jaes.2020.0070

Pörschmann, C., **Arend, J. M.**, Bau, D., & Lübeck, T. (2020). Comparison of Spherical Harmonics and Nearest-Neighbor based Interpolation of Head-Related Transfer Functions. In *Proc. of the AES International Conference on Audio for Virtual and Augmented Reality (AVAR), Redmond, WA, USA* (pp. 1–10).

Arend, J. M., & Pörschmann, C. (2019). Spatial upsampling of sparse head-related transfer function sets by directional equalization – Influence of the spherical sampling scheme. In *Proc. of the 23rd International Congress on Acoustics (ICA), Aachen, Germany* (pp. 2643–2650). https://doi.org/10.18154/RWTH-CONV-238939

Pörschmann, C., **Arend, J. M.**, & Brinkmann, F. (2019). Spatial upsampling of individual sparse head-related transfer function sets by directional equalization. In *Proc. of the 23rd International Congress on Acoustics (ICA), Aachen, Germany* (pp. 4870–4877). https://doi.org/10.18154/RWTH-CONV-239484

Pörschmann*, C., **Arend*, J. M.**, & Brinkmann, F. (2019). Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling. *IEEE/ACM Trans. Audio Speech Lang. Proc.*, *27*(6), 1060–1071. (*equal contributions). https://doi.org/10.1109/TASLP.2019.2908057

Pörschmann, C., & **Arend, J. M.** (2019). Obtaining Dense HRTF Sets from Sparse Measurements in Reverberant Environments. In *Proc. of the AES International Conference on Immersive and Interactive Audio (IIA), York, UK* (pp. 1–10).

Parametric Spatial Audio

Arend, J. M., Amengual Garí, S. V., Schissler, C., Klein, F., & Robinson, P. W. (2021). Six-Degrees-of-Freedom Parametric Spatial Audio Based on One Monaural Room Impulse Response. *J. Audio Eng. Soc.*, *69*(7/8), 557–575. https://doi.org/10.17743/jaes.2021.0009

Amengual Garí, S. V., **Arend, J. M.**, Calamia, P., & Robinson, P. W. (2020). Optimizations of the Spatial Decomposition Method for Binaural Reproduction. *J. Audio Eng. Soc.*, *68*(12), 959–976. https://doi.org/10.17743/jaes.2020.0063

Arend, J. M., Lübeck, T., & Pörschmann, C. (2019). A Reactive Virtual Acoustic Environment for Interactive Immersive Audio. In *Proc. of the AES International Conference on Immersive and Interactive Audio (IIA), York, UK* (pp. 1–10).

Pörschmann, C., Stade, P., & **Arend, J. M.** (2017). Binauralization of Omnidirectional Room Impulse Responses – Algorithm and Technical Evaluation. In *Proc. of the 20th International Conference on Digital Audio Effects (DAFx17), Edinburgh, UK* (pp. 345–352).

Auditory Distance Perception of Nearby Sound Sources

Arend*, J. M., Ramírez*, M., Liesefeld, H. R., & Pörschmann, C. (2021). Do near-field cues enhance the plausibility of non-individual binaural rendering in a dynamic multimodal virtual acoustic scene? *Acta Acust.*, *5*(55), 1–14. (*equal contributions). https://doi.org/10.1051/aacus/2021048

Arend, J. M., Liesefeld, H. R., & Pörschmann, C. (2021). On the influence of non-individual binaural cues and the impact of level normalization on auditory distance estimation of nearby sound sources. *Acta Acust.*, *5*(10), 1–21. https://doi.org/10.1051/aacus/2021001

Arend, J. M., Neidhardt, A., & Pörschmann, C. (2016). Measurement and Perceptual Evaluation of a Spherical Near-Field HRTF Set. In *Proc. of the 29th Tonmeistertagung - VDT International Convention, Cologne, Germany* (pp. 356–363).

Binaural Rendering of Spherical Microphone Array Data

Lübeck, T., **Arend, J. M.**, & Pörschmann, C. (2022). Binaural reproduction of dummy head and spherical microphone array data – A perceptual study on the minimum required spatial resolution. *J. Acoust. Soc. Am.*, *151*(1), 467–483. https://doi.org/10.1121/10.0009277

Arend*, J. M., Lübeck*, T., & Pörschmann*, C. (2021). Efficient binaural rendering of spherical microphone array data by linear filtering. *EURASIP J. Audio Speech Music Process.*, *2021*(37), 1–11. (*equal contributions). https://doi.org/10.1186/s13636-021-00224-5

Lübeck, T., Helmholz, H., **Arend, J. M.**, Pörschmann, C., & Ahrens, J. (2020). Perceptual Evaluation of Mitigation Approaches of Impairments due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data: Dry Acoustic Environments. In *Proc. of the 23rd International Conference on Digital Audio Effects (DAFx2020), Vienna, Austria* (pp. 250–257).

Lübeck, T., Helmholz, H., **Arend, J. M.**, Pörschmann, C., & Ahrens, J. (2020). Perceptual Evaluation of Mitigation Approaches of Impairments due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data. *J. Audio Eng. Soc.*, *68*(6), 428–440. https://doi.org/10.17743/jaes.2020.0038

Voice Directivity

Pörschmann, C., & **Arend, J. M.** (2021). Investigating phoneme-dependencies of spherical voice directivity patterns. *J. Acoust. Soc. Am.*, *149*(6), 4553–4564. https://doi.org/10.1121/10.0005401

Pörschmann, C., Lübeck, T., & **Arend, J. M.** (2020). Impact of face masks on voice radiation. *J. Acoust. Soc. Am.*, *148*(6), 3663–3670. https://doi.org/10.1121/10.0002853

Pörschmann, C., & **Arend, J. M.** (2020). A Method for Spatial Upsampling of Voice Directivity by Directional Equalization. *J. Audio Eng. Soc.*, *68*(9), 649–663. https://doi.org/10.17743/jaes.2020.0033

Contents

1 Introduction

Binaural rendering of spatial sound scenes is one key aspect of virtual acoustics. The binaural reproduction over headphones (or, less common, over loudspeakers) makes it possible to virtually place a listener in the acoustic scene, giving them the impression of being present and immersed in the virtual acoustic reality. Binaural technology is highly relevant in research as well as for consumer applications. It is used, among many others, in virtual or augmented reality (VR/AR), in acoustic simulations and auralizations, as well as in hearing science and neuroscience.

The fundamental idea of binaural technology is that the auditory experience of any sound field can be reproduced by accurately reconstructing its sound pressure at the eardrums of the two ears (Møller, 1992). The most straightforward approach would be to record the sound field using small microphones in the ear canal of a human listener or by using a dummy head, and then reproduce it (appropriately equalized) over headphones (Møller, 1992; Møller et al., 1995). A more common and flexible approach is binaural synthesis, where anechoic audio material gets convolved with a binaural impulse response. Such a binaural impulse response describes the acoustic transfer function between a sound source and a binaural receiver and can be either measured or modeled. In free-field conditions, the measuring or modeling process results in a head-related impulse response (HRIR) or its Fourier transform, the head-related transfer function (HRTF), whereas in reverberant conditions, the result is a binaural room impulse response (BRIR) or its Fourier transform, the binaural room transfer function (BRTF) (see, e.g., Wightman and Kistler (1989) or Algazi et al. (2001) for early work on HRTF measurements and modeling as well as Møller et al. (1996) or Kleiner et al. (1993) for early studies on BRIR measurements and modeling).

Binaural synthesis with a single binaural impulse response only allows the reproduction of one specific static source-receiver configuration, meaning that neither source nor listener movements are considered in the rendering. The advancement of this concept, called dynamic binaural synthesis, allows the binaural reproduction of a sound field for arbitrary source-receiver configurations. Dynamic binaural synthesis with 3 degrees of freedom (DoF) accounts for the listener's yaw, pitch, and roll head movements. In a virtual acoustic reality rendered in this way, the listener and sound source are fixed in position, with the listener free to move their head. Dynamic binaural synthesis with 6 DoF, however, also takes into account translational movements of the listener so that the listener can move freely in the virtual or acoustically augmented space. Likewise, depending on the implementation, the position and orientation of the sound source can be freely adjusted. This enables virtual acoustic scenarios in which the listener can, for example, freely place a (virtual) sound source in space and then walk towards or around it. To account for any listener or source motion in convolution-based dynamic binaural synthesis, binaural impulse responses corresponding to the particular source-receiver configuration (i.e., binaural impulse responses describing the acoustic transfer function for the listener's head orientation and position relative to the source orientation and position) are switched in real-time in the convolution engine. The required binaural impulse responses can either be provided as a database derived from prior measurements or modeling, or generated in real-time.

Considering that BRIRs are essential for creating reverberant virtual acoustic spaces, but extensive (individual) measurements are often not feasible, much research has been done on modeling (also called synthesizing) BRIRs for 3-DoF and 6-DoF applications. One way is BRIR synthesis based on room acoustic simulation, which mostly aims at the precise calculation of the spatial sound field based on a 3D room model, source and receiver models, and acoustic properties of the surfaces inside the room (Vorländer,

2008; Schröder, 2011; Schissler et al., 2017). Another method is parametric BRIR synthesis, which usually aims at perceptually motivated encoding and binaural decoding of the spatial sound field rather than a physically exact reproduction (Merimaa & Pulkki, 2005; Tervo et al., 2013; Pulkki et al., 2018). Using a spherical microphone array (SMA) is another flexible method to capture a spatial sound field and render it for a single listener over headphones. Based on spatial room impulse responses (SRIRs) measured with an SMA for one specific source-receiver configuration, BRIRs can be synthesized and then applied for dynamic binaural synthesis with 3 DoF (Avni et al., 2013; Bernschütz et al., 2014; Ahrens & Andersson, 2019). However, the major advantage of SMAs is that they can be used for 3-DoF real-time binaural rendering of a spatial sound scene, such as a musical performance in a concert hall. For this, the captured sound field is processed in real-time to generate binaural ear signals that, when presented over headphones, virtually place the listener at the position of the SMA in the room (Zotter & Frank, 2019; McCormack & Politis, 2019). Thus, starting from the fundamental idea, binaural technology has evolved tremendously, and nowadays, there are many different application-specific methods for the binaural rendering of virtual acoustic realities.

With the advent of VR, and especially in recent years with the growing interest in mobile AR devices and applications, the *efficiency* of binaural rendering has become increasingly important. Often, binaural rendering needs to be computationally lightweight, as the limited resources have to be shared with other computationally demanding components, such as visuals, sensors, and mapping, to name a few. At the same time, the rendering must be perceptually plausible and, in the case of AR applications, consistent with the real-world acoustics to create a coherent sound scene with real and virtual sound sources.

Optimizing efficiency can start from two different points. On the one hand, efficiency can be optimized by a purely technical approach, such as a smart implementation of an algorithm to save computational resources. On the other hand, efficiency optimization of a rendering implementation can be based on results of perceptual evaluations, meaning that a rendering algorithm is optimized (in an iterative process) concerning perceptual attributes. As an example, in development, psychoacoustic experiments are conducted to determine to what extent and how the virtual sound field can be simplified without decreasing perceptual plausibility. A simplified representation of the sound field, for instance, with a lower spatial resolution or fewer early reflections than in the physically correct sound field, usually directly impacts the required computational resources of the algorithm and thus leads to a more efficient implementation.

This thesis mainly focuses on *efficient binaural rendering* of virtual acoustic realities with 3 or 6 DoF. The thesis presents different methods applied in the context of headphone-based virtual acoustic realities and studies addressing recent psychoacoustic research questions in the field of binaural rendering. It discusses the suitability of the different methods for technically efficient rendering algorithms and how the perceptual results may influence the design and implementation of the examined methods and rendering algorithms. As such, this work addresses the state-of-the-art of headphone-based virtual acoustic realities from both a technical and perceptual perspective and provides fundamentals and ideas for future studies and implementations in the field of 3 and 6 DoF binaural rendering.

Section 1.1 first introduces the basics of spatial hearing and Section 1.2 reviews the current state of binaural rendering. These are extensive topics, but the thesis focuses on the essential key points necessary for the basic understanding of the work. Interested readers are kindly referred to further literature referenced in the corresponding chapters and sections to get a deeper insight into the topics. Section 1.3 provides an overview of the content of this thesis and explains how the research topics and publications included in this thesis relate to matters of technically or perceptually motivated efficient binaural rendering. Section 1.4 summarizes the original achievements of this work, and Section 1.5 provides future research perspectives. Chapters 2–6 contain the selected publications and thus form the main part of this thesis.

1.1 Spatial Hearing

Spatial hearing allows us to localize sound sources, distinguish between different sound sources in complex acoustic environments, and attend to sound sources of interest (Blauert, 1996). It aids orientation in space, enables communication in acoustically challenging environments, and allows us to hear out different instruments in an orchestra. Understanding the mechanisms of spatial hearing is essential for binaural rendering, which recreates spatial auditory cues to generate virtual acoustic realities.

The auditory system exploits binaural and monaural cues that originate from the head, to some extent from the torso, and the pinnae to localize the *direction* of a sound source. The binaural cues result from comparing both ear signals and mainly aim at localizing in the horizontal plane (left/right). They are a combination of interaural time differences (ITD), which mainly result from the distance between the ears, and interaural level differences (ILD), which mainly result from the acoustic shadowing of the head. The ITD and ILD cues operate in complementary frequency ranges. Below about 1.5 kHz, the ITDs provide stable localization cues, whereas the time differences become increasingly ambiguous at higher frequencies. Complementary, the ILDs are strongest above about 1.5 kHz because the wavelengths are small compared to the head and head shadowing increases (Stern et al., 2006). For broadband sound sources, low-frequency ITDs are the dominant cue for localization in the horizontal plane (Wightman & Kistler, 1992).

The monaural cues arise from spectral changes of the incoming sound caused by the head, the torso, and especially the pinnae. Sound is reflected in the pinna, creating direction-dependent resonances that alter the spectral content of the incoming sound. The resulting direction-dependent spectral profiles aid localization in the vertical plane (up/down). The (highly individual) monaural spectral cues allow localization of sound sources located in the median plane or on the cones of confusion, where almost no usable binaural cues are available. However, much of the front-back ambiguity in localizing sound sources that are in the median plane or on a cone of confusion can already be resolved by moving the head, creating binaural cues (Thurlow & Runge, 1967).

HRTFs describe the explained direction-dependent acoustic filtering of incoming sound by the listener's morphology. Thus, they contain all monaural and binaural auditory cues necessary for localizing the direction of a sound source. For this reason, HRTF filter sets, usually for a large number of directions, are essential for headphone-based reproduction of spatial sound fields.

In everyday listening situations, such as in rooms, the sounds we hear are usually a combination of direct sound and later arriving reflections from surfaces. With increasing delay after the direct sound, the reflections become denser in time, and their level decreases exponentially over time. These late dense reflections are grouped under the term reverberation. Whereas reverberation is usually spatially diffuse, the early reflections are strongly directional.

Reflections have numerous effects on sound perception, such as coloration or changes in the perceived width of a sound source (see, e.g., Toole (2008) for an extensive overview on the various effects of reflections). The accuracy in determining the direction of a sound source, however, is in most cases only slightly affected by reflections (Hartmann, 1983; Bech, 1998), mainly due to perceptual processes underlying the precedence effect (Litovsky et al., 1999; Brown et al., 2015). This effect has been studied extensively using the example of direct sound and a single ideal reflection reproduced by two loudspeakers. Depending on the delay time between direct sound and reflection, different percepts occur. A delay below 1 ms leads to the so-called summing localization, meaning the percept of a single fused sound image (phantom sound source) between the two loudspeakers and consequently to a shift in the perceived direction of the sound. For delay times above 1 ms, which are more common in real-life situations, the precedence effect occurs, and a fused sound image is perceived in the direction of the direct sound source, which means that the direct sound dominates the perceived direction. With longer delay times (approximately 5 ms for impulse

signals, 50 ms for speech, and 80 ms for music), the so-called echo threshold is exceeded, and the precedence effect no longer operates. In this case, two separate auditory events occur, that is, both the direct sound and the reflection are perceived separately coming from a specific direction (Blauert, 1996). Perceptual processes that can be attributed to the precedence effect thus allow to determine the direction of a sound source even in reflective environments. The precedence effect appears to be active also in real rooms with multiple reflections, meaning that the earlier arriving (direct) sound takes perceptual spatial precedence over the later arriving sound (Zahorik, 2021).

Reverberation, however, has a greater negative impact on correctly determining the direction of a sound source (Zahorik, 2021). The diffuse reverberation results in decorrelated signals with nearly equal energy at both ears. A relative increase in diffuse energy compared to direct sound and early reflections leads to less reliable ITD cues and much lower ILDs tending towards zero. The impaired binaural cues consequently reduce localization accuracy. As the reverberation level increases, the localizability of sound sources decreases, whereas other aspects of spatial perception, such as listener envelopment, become more pronounced. Both localizability and listener envelopment are closely linked to the interaural cross-correlation coefficient (IACC), which is a binaural measure describing the similarity (or coherence) of the ear signals (Okano et al., 1998). A decreasing IACC indicates a decrease in localizability, but an increase in listener envelopment and perceived spaciousness.

The primary auditory cues for localizing the *distance* of a sound source in the far field are intensity, direct-to-reverberant energy ratio (DRR), and spectrum (Zahorik et al., 2005; Kolarik et al., 2016). The relative intensity allows discriminating sounds at different distances, and the DRR, which decreases with increasing sound source distance, provides an absolute distance estimation cue independent of the sound source power in reverberant conditions. Hence, whereas reflected sound can impair accurately determining the direction of a sound source, it supports determining distance. Spectral cues for far-field sources appear only at distances greater than 15 m, due to high-frequency attenuation of the sound (Blauert, 1996).

Nearby sound sources at distances less than 1 m provide further specific distance cues. In particular, the ILDs exhibit significant distance-dependent changes for nearby lateral sound sources. In anechoic conditions and in the absence of the powerful intensity cue, specifically low-frequency ILDs ($f < 3$ kHz) appear to dominate distance perception of nearby lateral sources (Brungart, 1999). However, studies on intensity-independent distance perception of nearby sound sources in reverberant conditions yielded inconsistent results on whether the DRR cue masks the ILD cue or whether both cues support distance estimation, leaving the relative contribution of the different cues to intensity-independent distance perception unclear yet (Kopčo et al., 2020). Besides, nearby sound sources exhibit a distance-dependent low-pass filtering character due to a relative emphasis of low-frequency sound pressure caused by scattering at the head and torso, which might be a spectral distance estimation cue (Brungart & Rabinowitz, 1999). Last, acoustic parallax effects, which result in a lateral shift of some of the high-frequency features of the HRTF, may also be used for distance estimation of nearby sound sources (Zahorik et al., 2005; Kolarik et al., 2016).

1.2 Binaural Rendering of Virtual Acoustic Realities

There are various approaches for the binaural rendering of spatial sound scenes. For data-based dynamic binaural synthesis, which requires a set of precomputed binaural impulse responses, HRTFs or BRIRs can either be measured or generated using simulations. Furthermore, BRIRs can be generated by parametric synthesis. Depending on the method, the synthesis algorithm can be based only on a set of previously estimated parameters or on RIR measurements that form the basis for both the parameter estimation and the synthesis. Instead of precomputing BRIRs for data-based dynamic binaural synthesis, parametric rendering also allows the direct generation of binaural ear signals. Similarly, for SMAs, BRIRs can be generated

for data-based dynamic binaural synthesis based on measurements, whereas real-time processing of SMA signals allows directly generating binaural ear signals without synthesizing BRIRs in an intermediate step. The following section explains the different methods in further detail and outlines application scenarios for 3-DoF and 6-DoF binaural rendering.

Data-based dynamic binaural synthesis employing measured binaural impulse responses, either for individual subjects or with a dummy head, is still considered ground truth, and measurements are usually the reference for any simulation or modeling approach (Lindau, 2014; Brinkmann et al., 2017; Brinkmann, Aspöck, et al., 2019). Measuring binaural impulse responses in anechoic conditions yields HRTFs. Binaural rendering and BRIR synthesis usually require HRTF sets with a high spatial resolution, meaning a large number of HRTFs for different directions, and several approaches have been developed to acquire such high-resolution HRTF sets. Dummy head HRTFs are usually measured sequentially (Gardner & Keith, 1995; Bernschütz, 2013), which is rather time-consuming, whereas individual HRTFs are usually acquired using more complex procedures and equipment optimized for speed (Brinkmann, Dinakaran, et al., 2019; Richter, 2019). Acquiring (individual) full-spherical near-field HRTFs at different distances adds another layer of complexity (Xie et al., 2013).

Binaural impulse responses measured in reverberant environments – BRIRs – contain spatial auditory cues as well as information of the room, such as early reflections or reverberation. For 3-DoF dynamic binaural synthesis, BRIRs are measured for various head orientations for each fixed source-receiver configuration (Stade et al., 2012; Brinkmann et al., 2017). To reduce the measurement effort, often only horizontal head orientations are considered, covering the primary binaural cues. For 6-DoF rendering, measurements are performed for each sound source of interest for various head orientations at different positions in the room, resulting in a large number of BRIRs, depending on how finely the room is partitioned into grid cells (Werner et al., 2019). In the renderer, the BRIRs are then exchanged not only according to the relative head orientation to the virtual source but also according to the listener's position in the room, or more precisely, according to the grid cell where the listener is present. To keep the measurement effort feasible, Werner et al. (2018), for example, presented interpolation methods to generate BRIRs for the desired positions (grid cells) in the room based on measurements for different head orientations at only one or three positions in the room. Such interpolation methods have similarities to the fully parametric methods discussed later in this section.

Dynamic binaural synthesis with 3 DoF based on measured BRIRs has been extensively evaluated both technically as well as perceptually (see, e.g., Lindau (2014) for a detailed overview). Lindau and Weinzierl (2012) showed that dynamic binaural synthesis based on non-individual BRIRs can provide a perceptually plausible reproduction of a sound field. In this case, plausible means that, based on the listener's inner reference, the simulation was in accordance with their expectation towards the corresponding real sound field, which was evident in the listening experiment in that the subjects could not reliably detect whether a real or virtual sound source was presented. Later, Brinkmann et al. (2017) examined the authenticity of individual dynamic binaural synthesis, meaning whether the simulation was perceptually indistinguishable from the corresponding real sound field in a direct comparison task. The results showed that an authentic reproduction is only possible under certain conditions, such as speech in a reverberant room, whereas a noise stimulus or a dry room allowed the subjects to distinguish between simulation and the real sound field clearly. The findings suggest that under near-ideal technical conditions, authenticity depends in particular on how accurately spectral cues are reproduced and how reverberant the environment is.

Simulations are another way to obtain binaural impulse responses required for binaural rendering. Numerical simulation methods such as the boundary element method (BEM) are state-of-the-art to acquire HRTFs based on 3D head meshes (Ziegelwanger et al., 2015; Brinkmann, Dinakaran, et al., 2019). To

generate near-field HRTFs for any desired distance based on a set of far-field HRTFs, either range extrapolation (Duraiswami et al., 2004; Pollow et al., 2012) or distance variation functions (Kan et al., 2009; Spagnol et al., 2017) can be applied [1].

BRIRs can be created based on room acoustic simulations. Geometrical acoustic methods such as the image source method or ray tracing are the de-facto standard for the estimation of acoustic sound propagation in a virtual scene (Erraji et al., 2021). Often, the image source method and ray tracing are combined into a hybrid model, with the image source method used for spatially and temporally accurate simulation of direct sound and early reflections, and ray tracing used for simulation of the scattered sound field and late reverberation (Vorländer, 2008; Schröder, 2011). Using an HRTF set for directional filtering of the sound field components, BRIRs can then be synthesized for arbitrary head orientations and source-receiver configuration based on the room acoustic simulation. Accurate real-time simulation with a high spatial resolution and subsequent real-time binaural rendering for VR/AR applications with 3 or 6 DoF is highly computationally demanding (Schröder, 2011). Therefore, in consumer applications, for instance, the spatial resolution is adaptively adjusted based on perceptual metrics, which saves computation and is thus faster (Schissler et al., 2017).

Room acoustic simulations have also been evaluated in detail, in particular from a technical point of view (see, e.g., Vorländer (2008) or Savioja and Svensson (2015) for comprehensive overviews). Brinkmann, Aspöck, et al. (2019) further provided a detailed perceptual evaluation comparing 3-DoF binaural rendering using simulated BRIRs with renderings using measured BRIRs. Most of the investigated simulation algorithms were perceptually plausible, meaning that the subjects could mostly not detect whether the rendering was based on simulated or measured BRIRs. However, none of the simulation algorithms provided an authentic reproduction, meaning that in direct comparisons, the subjects could always hear a difference between simulated and measured BRIRs. Again, spectral differences as well as deviations in the perceived source position, likely occurring because of various inaccuracies in the simulations, were the reason why the subjects could perceive differences.

A currently popular topic is *parametric binaural rendering*, primarily because it promises perceptually plausible spatial audio reproduction at relatively low computational cost. The basic idea of this approach is to describe the properties of the sound field using parameters, such as time of arrival (TOA) or direction of arrival (DOA) of the direct sound and early reflections or characteristics of the reverberation such as diffuseness or reverberation time (Pulkki et al., 2018). The parameter estimation, also called encoding, is usually perceptually motivated and is mostly based on (S)RIR measurements obtained in the corresponding room. For binaural rendering, also called decoding, either BRIRs for data-based dynamic binaural synthesis or binaural ear signals are generated based on the parameters. The decoding is also usually perceptually motivated and scalable to some degree, meaning that the decoding accuracy can often be adjusted, for example, by changing the number of dynamic early reflections or the spatial resolution. This allows for adaption to various technical conditions and available computational resources. Moreover, the parameters can be easily adjusted before decoding to represent, for example, different room acoustic situations or source-receiver configuration, making parametric rendering a highly flexible method for spatial audio reproduction.

Various methods for parameter estimation and parametric rendering of sound fields have been presented, such as spatial impulse response rendering (SIRR) by Merimaa and Pulkki (2005), directional audio coding (DirAC) by Pulkki (2007), or the spatial decomposition method (SDM) by Tervo et al. (2013), to name a few popular approaches (see also Pulkki et al. (2018) for a comprehensive overview of parametric methods). The methods generally en- and decode the sound field for one specific listener position in space

[1] Even though these methods are more related to HRTF synthesis/processing than to simulation, they are mentioned here for completeness.

and, when employing headphone-based reproduction of the sound field, allow 3-DoF binaural rendering. The BRIR synthesis underlying most approaches either uses signal components of the measured RIRs in addition to the parametric description (Merimaa & Pulkki, 2005; Pulkki, 2007; Tervo et al., 2013) or relies solely on the determined parameters (Stade et al., 2017; Coleman et al., 2017; Brinkmann et al., 2020). Similar to simulations, directional components, such as direct sound and early reflections, are generated employing directional filtering with HRTFs, making high-resolution HRTF sets essential for parametric binaural rendering. Recent work extended existing methods to render the sound field for 6 DoF, meaning for arbitrary head orientations and listener positions in the room. To obtain a description of the sound field for any position, the methods either extrapolate the parametric sound field based on one single SRIR measurement (Pihlajamäki & Pulkki, 2015) or interpolate between multiple SRIR measurements distributed in the room (Müller & Zotter, 2020).

In most cases, studies on new parametric methods include a technical and perceptual evaluation, resulting in many different perceptual results from often differing test designs. In general, however, synthesized BRIRs are usually compared with measured BRIRs, or in fewer cases, with simulated BRIRs. In technical evaluations, BRIRs are often compared regarding various acoustic parameters such as spectrum, reverberation time, or IACC. To estimate potential perceptual influences of the deviations, the differences are often related to the respective parameters' just-noticeable differences (JNDs). The results show that parametric synthesis can produce BRIRs with differences below the JNDs, indicating that the synthesis methods are technically accurate (Zaunschirm, Frank, & Zotter, 2018; Amengual Garí et al., 2019). Consequently, researchers presented satisfactory experimental results indicating high perceptual similarity of 3-DoF binaural rendering using synthesized or measured BRIRs (Stade et al., 2017; Zaunschirm, Frank, & Zotter, 2018; Ahrens, 2019). However, in direct comparisons, synthesized and measured BRIRs are usually distinguishable due to minor spectral differences or slightly different spatial properties. Authenticity and plausibility studies comparing parametric binaural rendering of virtual sound sources to real loudspeaker sources are rare. Amengual Garí et al. (2019) reported that listeners in a direct comparison task could reliably detect the loudspeaker primarily based on spectral differences, suggesting that parametric 3-DoF binaural rendering does not yet provide an authentic reproduction. In the same study, however, the authors showed that the presented parametric rendering method provides a perceptually plausible reproduction in critical comparisons with a real loudspeaker source. Perceptual assessments regarding the perceived plausibility of 6-DoF parametric rendering, in which the listener is free to move around the room, have to the best of the author's knowledge not been published before the work on 6-DoF parametric rendering included in this thesis.

Also currently very popular is *binaural rendering of SMA data*. Based on SMA measurements, BRIRs can be synthesized that ideally (according to the mathematical theory) match measured individual or dummy head BRIRs. Furthermore, the use of SMAs allows the dynamic real-time rendering of spatial sound scenes, which is particularly interesting, for example, for live concert streaming or VR teleconferencing. To generate BRIRs or binaural ear signals, the sound field captured with the SMA is first spatially encoded by transforming it into the spherical harmonics (SH) domain using the discrete SH transform, where the spatial accuracy of this transform is defined by the SH order N (Rafaely, 2015). The resulting SH signals are then processed with radial filters. These are array-specific filter functions that compensate for the spatial extent and, in the case of a rigid sphere array, the scattering properties of the array body (Bernschütz et al., 2014; Ahrens & Andersson, 2019). Decoding these SH signals (also called Ambisonics signals) for 3-DoF binaural rendering can be done in several ways. A classical method is the virtual loudspeaker approach, where spatially uniformly distributed plane waves are generated by applying the inverse SH transform to the SH signals. The plane waves are then weighted with HRTFs for the corresponding directions and summed up, resulting in the binaural signal (McKeag & McGrath, 1996; Jot et al., 1999; Bernschütz et al.,

2014; Zotter & Frank, 2019). More recent methods perform binaural rendering directly in the SH domain. In this case, the full-spherical HRTF set is transformed to the SH domain and multiplied there with the SH signals of the array. Summing up the resulting directionally weighted SH signals yields the binaural signal (Helmholz et al., 2019; Zotter & Frank, 2019). From a mathematical point of view, both methods lead to the same result, but depending on the practical implementation, the resulting binaural signals may differ. Binaural rendering with 6 DoF based on SMA data and SH processing is currently increasingly researched. The approaches presented so far usually distribute several SMAs in space and then interpolate between the different local sound field descriptions (Patricio et al., 2019) or calculate the sound field over the entire area covered by the SMAs (Kaneko & Duraiswami, 2021).

Apart from the SH-based approach, there are also beamforming-based approaches for generating binaural signals from data of spherical or even planar microphone arrays (Madmoni et al., 2020; Fallahi et al., 2021). However, the focus of this work is on SH-based methods using SMAs, so the reader interested in beamforming is referred to other literature. Furthermore, the described SMA processing is in parts similar to some parametric methods, especially because parametric methods often also use SMAs and the SH transform for spatial encoding of the sound field. However, in the further processing of the SH signals as well as in the decoding, the described SMA processing differs from the parametric methods, as the SMA processing is based on a closed-form mathematical solution physically accurately describing the sound field, whereas the parametric decoding usually represents the sound field in a rather simplified way and does not necessarily attempt to reconstruct it precisely.

When measuring or recording a sound field using real SMAs with a limited number of microphones, spatial undersampling errors in the form of spatial aliasing and SH order truncation errors occur caused by the spatial discretization of the sound field. This leads to spatial and spectral distortions in the captured sound field, especially towards higher frequencies above the so-called spatial-aliasing frequency (Rafaely, 2015; Bernschütz, 2016). Furthermore, for binaural decoding, a low-order sound field (usually $N \leq 7$) must be combined with a high-order HRTF set (usually $N \geq 35$). The simplest and at the same time worst approach would be to truncate the SH order of the HRTF set to match the order of the sound field, resulting in further spatial and spectral distortions. To mitigate the artifacts that arise during capturing and decoding, various pre- and postprocessing methods have been developed, such as max-$\mathbf{r}_\mathrm{E}$ weighting, spatial resampling, spherical head filters, or MagLS, which are commonly applied nowadays for binaural rendering of SMA data (Zotter & Frank, 2012; Bernschütz et al., 2014; Ben-Hur et al., 2017; Zotter & Frank, 2019). For the technical evaluation of such rendering chains, simulated sound fields, sometimes reduced to a single plane wave, are often compared to the reconstructed sound field acquired with an SMA to estimate the influence of the SMA and the binaural decoding (Bernschütz, 2016; Ben-Hur et al., 2018). In addition, BRIRs synthesized from SMA measurements or SMA simulations are usually compared to measured or simulated BRIRs in terms of general spectral differences, deviations in monaural and binaural cues, and variances in spatial parameters (Ben-Hur et al., 2017; Zaunschirm, Schoerkhuber, & Hoeldrich, 2018; Engel et al., 2021). In particular, due to the low SH order of real SMAs, the differences are mostly above the respective JNDs despite state-of-the-art binaural decoding, where most likely coloration artifacts have the greatest perceptual influence. However, listening experiments showed satisfying results regarding the perceived quality of 3-DoF binaural rendering based on SMA data, which naturally improves with increasing SH order of the SMA. Thus, for SH orders $N \geq 7$, the perceptual differences between synthesized and measured BRIRs become significantly smaller, leading to similar quality ratings (Bernschütz et al., 2014; Zaunschirm, Schoerkhuber, & Hoeldrich, 2018; Ahrens & Andersson, 2019). Perceptual assessments regarding the perceived authenticity or plausibility of 3-DoF or even 6-DoF binaural rendering of SMA data compared to real sound sources have not yet been published to the best of the author's knowledge.

With the recent increased focus on 6-DoF binaural rendering, *position-dynamic* sound field changes occurring when a listener walks through the room have become a growing research topic. One position-dynamic factor that is particularly important in 6-DoF rendering is the sound source directivity, as it significantly affects the direct sound, the reflected sound, and the associated DRR. Incorporating directivity is essential for a plausible representation of scenes where a listener moves in front of or even walks around a sound source. That said, research is currently highly interested in measuring human voice directivity, estimating its dynamic behavior for fluent speech, and integrating it in 6-DoF binaural rendering chains, as it is crucial for plausible auralizations of singers or binaural rendering of virtual avatars' speech in VR/AR scenarios, to name a few examples. Furthermore, position-dynamic distance perception and accurate adjustment of distance cues are essential factors in 6-DoF reproduction. Particularly as listeners in 6-DoF environments can dynamically change the distance to a sound source and get very close or, in the case of handheld sound sources, even hold it very close to the head, binaural near-field rendering and the related near-field HRTFs regain importance. Besides these more technical factors, the position-dynamic sound field changes also lead to new psychoacoustic research. For example, most studies assessed perceptual parameters under unimodal conditions with 3-DoF rendering. However, in a multimodal dynamic 6-DoF environment, perceptual parameters, such as JNDs, can shift, and position-dynamic effects that have a strong perceptual impact can occur. At the same time, common test designs for 3-DoF environments have to be adapted to the changed conditions of a 6-DoF experiment, especially as subjects are no longer in one location but can walk freely through the (virtual) room.

1.3 Technical and Perceptual Concepts for Efficient Rendering

The present thesis covers a broad range of topics in the field of binaural rendering and addresses a variety of different state-of-the-art rendering methods as well as recent technical and psychoacoustic research questions in the respective areas. Despite this thematic diversity, all included studies have in common that they directly contribute to advancing *efficient* binaural rendering. The thesis considers efficiency from a technical and perceptual point of view, even though these are two interdependent aspects, as a perceptually efficient binaural rendering usually leads to a technically more efficient method. While all applications employing binaural technology generally benefit from efficient rendering, the rapidly increasing research and development of mobile (consumer) VR/AR applications and systems with relatively low computing power has further increased the focus on technical and perceptual concepts for efficient rendering. The following paragraphs provide a brief overview of Chapters 2–6, which contain the selected publications for this thesis, organized by research topic. Thereby, each paragraph elaborates how the chapter's content relates to efficiency, the technical and perceptual concepts presented for efficient binaural rendering, and the chapter's contribution to the main objectives of this thesis.

Chapter 2 deals with HRTF processing and interpolation, focusing in particular on spatial upsampling of sparse HRTF sets using the proposed SUpDEq method (SUpDEq - Spatial Upsampling by Directional Equalization), which is a pre- and postprocessing method for SH-based interpolation of HRTFs. For one thing, the presented method aims to efficiently represent the HRTF in the SH domain. Furthermore, the basic idea of the upsampling method is to generate a full-spherical HRTF set with a high spatial resolution based on only a small number of measurements. Applying the method significantly reduces the effort of measuring individual HRTFs and thus makes acquiring individual HRTFs much more efficient. As detailed in one of the studies presented in this chapter, the developed SUpDEq method can also be successfully applied to measurements in reverberant environments. Combining the proposed processing with a simplified measurement setup, consumer HRTF measurement systems can be developed. The newest study in this chapter assesses various SH-based interpolation methods recently proposed in literature, re-

vealing similar performance across all examined methods. Moreover, the study determined the minimum SH order required for perceptually transparent interpolation using the proposed SUpDEq method or unprocessed SH interpolation in a listening experiment. The results show that the required SH order and thus the required number of HRTF directions significantly decreases when using the SUpDEq method for interpolation. Depending on the application and the source signal, an SH order of $N \approx 7$, corresponding to only 64 spatially uniformly distributed HRTF directions, is sufficient for transparent SH interpolation when using the SUpDEq method.

Chapter 3 addresses parametric encoding and binaural decoding of parametrically described spatial sound fields. As parametric rendering is an efficient way to achieve a high-quality spatial audio reproduction that is scalable to the particular technical conditions, the presented methods are highly relevant, especially for 6-DoF AR applications with limited resources. Through perceptually motivated encoding, the presented methods reduce the sound field to its perceptually relevant components and describe it using a small number of acoustic parameters. The parametric description of the sound field, which forms the basis for parametric BRIR synthesis or parametric real-time rendering, allows flexible adjustments of the virtual acoustic scene by changing the parameters. The chapter presents various parametric encoding and decoding approaches for specific use-cases, such as 3-DoF and 6-DoF binaural rendering for VR/AR applications or binaural reproduction of self-generated sounds in a virtual acoustic reality. Furthermore, one of the studies presented in this chapter examines the parametric encoding of the SDM method in detail and presents (perceptually motivated) optimizations of the method for binaural decoding. Two studies in this chapter present plausibility experiments that compare binaural rendering with real loudspeaker sources. The plausibility evaluation of the SDM optimizations focuses on the spatial quantization of the reflections' DOAs. Subjects perceived 3-DoF binaural renderings with 14 spatially uniformly distributed DOAs as plausible as the real loudspeaker, indicating that the spatial resolution of reflected sound can be drastically reduced without perceptual degradations. The newest study in this chapter presents a pilot study assessing the plausibility of 6-DoF parametric binaural rendering based on one monaural RIR used for en- and decoding. The results show that the proposed system can provide a plausible binaural reproduction, but it also revealed challenges of 6-DoF rendering requiring further research.

Chapter 4 discusses auditory distance perception of nearby sound sources in virtual acoustic realities and the measurement or synthesis of near-field HRTFs as required for binaural reproduction of nearby sound sources. Binaural near-field rendering is particularly relevant for spatial auditory displays and has gained further importance by the increasing interest in 6-DoF applications. The presented studies aim to determine how the various auditory near-field cues contribute to distance perception and whether near-field cues enhance the plausibility of binaural rendering in dynamic multimodal virtual acoustic scenes. The latter is particularly important for complex real-time applications with limited computing resources, such as mobile AR applications. Here, it is crucial to know whether the additional computational cost of including near-field cues is worthwhile in terms of plausibility and overall reproduction quality or whether nearby sound sources can be simply rendered by using intensity-scaled far-field HRTFs, which is computationally much more efficient. One of the studies presented in this chapter presents three experiments examining how binaural cues contribute to distance estimation of nearby sound sources. In this context, the study shows that the normalization methods often used in such experiments to suppress intensity cues actually leave (incorrect) intensity cues, which might in parts explain conflicting findings regarding the effectiveness of binaural cues for relative distance estimation in the literature. The newest study in this chapter presents two experiments assessing the plausibility of near-field rendering based on either synthesized near-field HRTFs or intensity-scaled far-field HRTFs in a dynamic multimodal virtual acoustic scene. Participants controlled the virtual sound source position by moving a small handheld loudspeaker along

a prescribed trajectory, which provided visual and proprioceptive cues in addition to the auditory cues. The results of both experiments show no evidence that near-field cues enhance the plausibility of 6-DoF binaural rendering of nearby anechoic sound sources, indicating that, at least in terms of plausibility, the additional effort of including near-field cues may not be worthwhile.

Chapter 5 concerns binaural rendering of SMA data. Two studies presented in this chapter describe listening experiments on different approaches to mitigate spatial aliasing and SH order truncation errors that occur when using real SMAs with a limited number of microphones. Subjects rated the perceived qualitative differences between 3-DoF binaural rendering using measured dummy head BRIRs or synthesized BRIRs from SMA data with and without mitigation approaches. The results show that most mitigation approaches perceptually improve the reproduction, although audible differences to the reference remain. Furthermore, the experiments revealed that the magnitude of improvement is comparable across approaches and acoustic environments. Another study in this chapter presents a linear filtering approach for binaural reproduction of SMA data that is computationally more efficient than the conventional SH-based approaches. The study reviews conventional approaches for real-time binaural rendering of SMA data and shows that any rendering chain, including popular mitigation approaches, can be represented as a set of precomputed finite impulse response filters. These filters are then applied to the SMA signals in real-time using fast convolution to produce the binaural signals. The study includes working examples and sample calculations on computational complexity and memory requirements for the technical evaluation of the proposed method. This chapter's most recent perceptual study presents two listening experiments to determine the minimum SH order for direct sound, early reflections, and reverberation of dummy head or SMA measurements required to generate 3-DoF binaural renderings perceptually indistinguishable from a high-resolution reference. The listening experiments revealed that dummy head BRIRs require a significantly lower SH order than SMA BRIRs, mainly because the sound field is error-free when measuring discrete BRIRs, whereas SMA captures always suffer from spatial undersampling errors. Furthermore, the experiments revealed that much lower SH orders are necessary for early reflections and reverberation than for direct sound, indicating that the spatial resolution of reflected sound can be significantly decreased without perceptual impairments. The results of this study form the basis for developing perceptually efficient algorithms for binaural rendering of SMA signals.

Chapter 6 reports studies on human voice directivity, a topic of particular importance in 6-DoF virtual acoustic realities for the accurate reproduction of speech. The long-term goal of the presented studies is to determine how accurately, that is, with which temporal and spatial resolution, the dynamic voice directivity must be rendered. First, however, the dynamic directivity must be acquired, usually done using surrounding spherical microphone arrays. One of the studies in this chapter presents a method for spatial upsampling of such simultaneous measurements, which is an efficient approach for obtaining high-resolution directivity data from a small number of measurements. The proposed approach is based on the SUpDEq method and combines pre- and postprocessing (directional equalization and de-equalization) with SH-based interpolation. Based on measurements of a dummy head with an integrated mouth simulator, the study compares the spatial upsampling approach to reference measurements on a dense grid. The results show that the method significantly decreases spatial undersampling errors and thus allows to determine a meaningful high-resolution voice directivity from sparse measurements. The newest study in this chapter applied the previously evaluated spatial upsampling method to voice directivity measurements of human subjects. The work examines the estimated spherical voice directivity patterns regarding different phonemes and phoneme-dependent variations and reveals significant phoneme-dependent differences. In future perceptual studies, the phoneme-dependent voice directivity data will be used for binaural

renderings to determine whether voice directivity must be reproduced in such detail or whether a computationally more efficient directivity model is perceptually sufficient.

1.4 Original Achievements

The studies included in this thesis address various technical and psychoacoustic aspects of binaural rendering. Their results provide concepts, methods, and design parameters for the implementation of technically and perceptually efficient rendering. The following paragraphs briefly summarize the original achievements of the studies included in this thesis, organized by research topic.

Head-Related Transfer Function Processing and Interpolation The author's studies in this field mainly concern the SUpDEq method, a novel approach for SH-based HRTF interpolation and spatial upsampling of sparse HRTF sets. The various studies evaluated the method for individual and non-individual HRTFs, different spherical sampling schemes, and under anechoic and reverberant conditions. The studies showed that the method is robust and provides comparable results under all conditions, significantly outperforming unprocessed SH interpolation in all cases. Furthermore, the author presented the first study technically comparing four state-of-the-art SH-based HRTF interpolation approaches (including the SUpDEq method), all of which time-align the HRTFs in different ways before the SH transform. The study explained the different methods in the same mathematical framework and clarified the similarities and differences of the approaches. As an important outcome, the study revealed very similar performance of the methods, with notable differences only at low SH orders and contralateral HRTFs. Accordingly, depending on the application, the choice of method may be driven more by factors other than performance, such as computational demands or implementation effort. Furthermore, the work presented the first psychoacoustic study on the minimum SH order required for perceptually transparent interpolation when using the SUpDEq method or common unprocessed SH interpolation. Based on the similarity between the different methods found in the technical evaluation, it appears reasonable that the perceptual results obtained for the SUpDEq method are approximately valid for the other methods as well. The estimated thresholds provide the starting point for further quality-based listening experiments.

Parametric Spatial Audio The author presented a unique method for 6-DoF parametric spatial audio rendering based on one monaural RIR. The work detailed the entire pipeline to derive a spatial parametric description of the sound field from a monaural RIR and generate synthetic BRIRs for any desired head orientation and position in the room, as well as a purpose-built real-time framework that can be used to perform psychoacoustic experiments for research on AR. The study further presented the first 6-DoF plausibility experiment in an AR scenario comparing four real loudspeaker sources with their virtual counterparts, suggesting that 6-DoF parametric binaural rendering based on one monaural RIR can provide perceptually plausible results. Other studies by the author presented state-of-the-art algorithms for parametric BRIR synthesis based solely on previously estimated parameters. The author used the proposed binaural rendering methods also for a reactive virtual acoustic environment, a system for directional capturing and binaural reproduction of self-generated sound that the author has designed and implemented to study, for example, the influence of room acoustics on musicians' performance or the effect of self-generated sounds such as speech on presence. Furthermore, the author presented optimizations of the SDM method for binaural rendering. The study evaluated the encoding of SDM in detail, providing important insights and principles for its application, and proposed optimizations technically and perceptually improving the binaural decoding of SDM measurements. Furthermore, the plausibility experiment presented in the SDM study provided new insights on the required spatial resolution of reflected sound when using parametric binaural rendering.

Auditory Distance Perception of Nearby Sound Sources The author detailed the acquisition of a near-field HRTF database of the Neumann KU100 dummy head, containing circular and full-spherical measurements at sound source distances between 0.25 m and 1.50 m. Employing those measurements, the author conducted a study on the influence of binaural cues and the impact of level normalization on auditory distance estimation of nearby sound sources. The author's article contains a detailed review of studies in this field, revealing the long-standing controversy regarding whether humans can utilize binaural cues for distance estimation of nearby sound sources. The series of three listening experiments showed that experimental procedures applying established normalization methods for compensating intensity cues are not suitable for correctly investigating the influence of binaural cues on relative distance estimation, mainly because salient intensity cues remain dominating distance judgments. The results, however, revealed that those drawbacks of the test method might partly explain the conflicting findings in literature regarding the effectiveness of binaural cues for relative distance estimation. The author further conducted the first study on perceptual plausibility of near-field rendering by running two psychoacoustic experiments in a 6-DoF multimodal AR experience. The results of both experiments provide no evidence that near-field cues enhance the plausibility of binaural rendering, suggesting that including near-field cues or near-field HRTF synthesis in binaural rendering may not be necessary in terms of plausibility and reproduction quality of multimodal scenes.

Binaural Rendering of Spherical Microphone Array Data The author presented two studies that, for the first time, technically and perceptually compared different state-of-the-art approaches to mitigate spatial undersampling errors in binaural rendering of SMA data. The studies clarify the different impacts of spatial aliasing and SH order truncation and explain at which point in the rendering chain the different mitigation approaches operate. Importantly for the scientific community, the perceptual study showed that most methods perform similarly well and provided better perceptual results than binaural rendering without mitigation processing. Another study by the author proposed a new method for more efficient binaural rendering of SMA data using linear filtering. The proposed method relies on a precomputed filter set and a standard dynamic binaural renderer, greatly simplifying the rendering chain. Besides, the study provides a sound overview of the current state of the art in binaural rendering of SMA data and might thus offer a good introduction to this field. The author further presented the first perceptual study determining the minimum SH order of dummy head or SMA measurements for the three basic time intervals of the sound field – direct sound, early reflections, and reverberation – required for perceptually transparent 3-DoF binaural rendering in comparison to a high-resolution reference. Vital to the community, the study provides a fundamental understanding of the different errors in SH interpolation of BRIRs and binaural rendering of SMA data. The results of the listening experiments, which revealed lower minimum required SH orders for dummy head BRIRs than for SMA BRIRs and a general decrease in the required SH order for later parts of the sound field, provide insights into basic spatial hearing processes, a better understanding of SH processing, and important guidelines for the development of more efficient rendering algorithms in the future.

Voice Directivity The author presented a study describing a novel method for spatial upsampling of voice directivity measured with a surrounding spherical microphone array. The study successfully adapts the SUpDEq method, initially developed for upsampling of HRTF sets, to voice directivity measurements and shows that applying the proposed upsampling method significantly decreases spatial undersampling errors and allows determining meaningful high-resolution voice directivity patterns from sparse measurements. Another study presented by the author investigated for the first time - due to current events at that time - the impact of face masks on voice radiation. The results showed how different masks affect transmission loss and voice directivity. Although not directly relevant to practical applications, it had a far-reaching impact on the entire scientific community, increasingly studying the various influences of face masks at the

time of the study. The author further presented the first study on the acquisition and evaluation of full-spherical voice directivity measurements for different phonemes. The study successfully applied the previously introduced SUpDEq method to human voice directivity measurements with a surrounding spherical microphone array. The work showed significant differences in the directivity of the phonemes, providing new general insights into articulation-dependent aspects of human voice directivity. The results can contribute to models of human voice production and be used for VR/AR applications and room simulations to integrate accurate radiation patterns in sound field rendering and computation.

Other significant achievements include data and software toolboxes, which the author usually obtained as part of the studies, and published open source to share the data and tools with the community and support open research. The chapter on open software and data contributions at the end of this thesis lists the author's most relevant open source contributions.

1.5 Future Research Perspectives

There is a large body of research on binaural technology and rendering and many future research perspectives accordingly. Explaining current and future research questions in all areas of binaural rendering and spatial audio is beyond the scope of this section. For this reason, the following paragraphs briefly summarize some of the author's research ideas that may be relevant in the future to the research topics discussed in this thesis.

Research on HRTF processing and SH-based interpolation aims at further reducing the required minimum number of HRTF directions for perceptually transparent interpolation. Current SH-based interpolation methods yield the largest errors at the contralateral ear, primarily because the HRTF at the contralateral side, even after preprocessing (i.e., time-alignment) of the HRTF, still exhibits high spatial complexity due to distinct magnitude interference patterns that change substantially with even small changes in source position. As the listening experiment presented in this thesis showed that these distortions at the contralateral ear are audible, future research should develop improved interpolation algorithms that further reduce the errors at the contralateral side.

It is also important to investigate the combination of different interpolation methods, such as the combination of natural neighbor and SH-based interpolation of time-aligned HRTFs. The procedure can be helpful, for example, when dealing with arbitrary, non-ideal spherical sampling grids, as is often the case with self-guided HRTF measurement systems. Transforming such HRTF sets to the SH domain for interpolation usually results in strong interpolation errors. Instead of stabilizing the SH transform, for example, by applying Tikhonov regularization, which can also be effective, it seems to be a good approach to first generate stable full-spherical sampling grids using, for instance, natural neighbor interpolation and then apply spatial upsampling to those (sparse) full-spherical HRTFs sets using SH-based methods. However, the parameterization and performance of combined methods require further extensive research.

For comparability, perceptual evaluations of future improved interpolation methods should also aim to determine the required minimum number of HRTF directions using a similar test design to that presented in this thesis. Furthermore, future perceptual studies should focus on quality estimation of binaural renderings of more application-oriented reverberant scenes. For example, virtual acoustic realities rendered using different spatially upsampled HRTFs or dense reference HRTF sets could be compared using the SAQI paradigm (Lindau et al., 2014). Such studies would yield a valid estimate of the required minimum number of (individual) HRTFs and interpolation methods suited for consumer applications.

Much of future research on parametric spatial audio will address 6-DoF rendering and related encoding, interpolation, extrapolation, and decoding approaches. Real-time estimation of room acoustic parameters for parametric binaural rendering in AR applications and devices remains a major challenge and requires further research. Determining parameters based on the multimodal input, blind system identification for estimating the reverberation time, or machine learning based approaches for estimating reverberation parameters or absorption parameters of surfaces based on images are just a few examples that will be intensively researched in the future. Moreover, many questions remain about how best to extrapolate parameters determined at one position in a room to other positions in the room, especially when no geometric information about the room is available or can be determined. However, thinking about future AR devices with numerous different sensors, research is required on how to combine the different inputs beneficially, for example, to better extrapolate room acoustic parameters based on the geometrical information provided by a depth camera. Besides, parametric spatial audio offers many other research questions, such as how to integrate sound source directivity correctly, how to handle and efficiently integrate directional reverberation, how to encode and decode coupled rooms with unusual reverberation tails, how to handle drastic parameter changes when a user walks through acoustically highly different rooms in an AR scenario, or how to decode and extrapolate parameters of acoustically inhomogeneous rooms.

The trend towards (parametric) 6-DoF rendering also raises many new perceptual research questions. Future listening experiments should examine perceptually salient position-dynamic effects typical for 6-DoF environments that must be reproduced accurately for plausible rendering. In this context, future psychoacoustic studies should determine the required accuracy of parametric rendering, such as the number of dynamically reproduced early reflections, and investigate to what extent the determined number of early reflections is position-dependent. Similarly, the question arises how exactly TOAs and DOAs of early reflections have to be extrapolated and decoded, or whether it is even necessary from a perceptual point of view to dynamically adjust the early reflections as a listener walks through the room. Such questions could be investigated, for example, with experiments comparing the perceived plausibility of various rendering methods. Also exciting would be to investigate the perceptual impact of directional (i.e., not fully diffuse) reverberation and examine whether listeners can perceive slight changes in the directivity of the reverberation as they walk through a room. Regardless of these more specific research questions, many of the conventional test designs for 3-DoF experiments must first be adapted to 6-DoF conditions. For example, the question arises on how to achieve reproducible results when participants are free to move around the room, generating different cues and individual test runs. Hence, for any 6-DoF experiment, a trade-off must be made between additional inaccuracies in the data when subjects are free to move or more reproducible results when walking paths and movements are restricted and predetermined but therefore somewhat unnatural.

In research on auditory distance perception of nearby sound sources, further studies are needed to clarify the relative contribution of the DRR and ILD cues to intensity-independent distance perception. In addition, new test paradigms are required to investigate distance perception of nearby sound sources under natural conditions (i.e., including the intensity cue), to determine the influence of the ILD cue, and to examine the interaction of the DRR and ILD cues in the presence of the dominant intensity cue. For example, experiments to estimate the JND of auditory distance estimation for nearby sound sources with and without the distinct low-frequency ILD cues could reveal whether the ILD cues provide a benefit and contribute to distance estimation. Such experiments would be another step towards solving the long-standing controversy regarding whether humans can actually utilize binaural cues for distance estimation of nearby sound sources. Moreover, future research may clarify the currently conflicting results on whether individual HRTFs improve auditory distance perception over non-individual HRTFs. Besides, in the context of future VR and AR systems and applications, psychoacoustic experiments on distance perception

of (nearby) sound sources should be conducted in more realistic multimodal conditions involving various cues (visual, proprioceptive, auditory), as this represents the way we perceive our environment much better than unimodal experimental setups.

Research on binaural rendering of SMA data will further focus on improving and developing methods to mitigate spatial undersampling errors. However, most approaches attempt to reduce the spatial undersampling errors after the SH transform of the SMA signals. Future research should aim for novel methods to preprocess the sparse SMA data before SH transform, as done in SH-based HRTF interpolation, to reduce the errors occurring when transforming sparse SMA data to the SH domain. In addition, further research is needed on how to integrate knowledge on the required spatial resolution of different sound field components, as determined in listening experiments, into real-time rendering chains to make them more efficient. A more recent trend is beamforming-based binaural rendering, as it is not limited to spherical arrays but can generally be applied to any array geometry. Future research should provide a comprehensive background on beamforming-based methods and explain and compare the SH-based and beamforming-based approaches in a similar mathematical framework. Besides, further research is required on 6-DoF binaural rendering based on SMA, for example, to clarify how to best extrapolate a sound field captured with an SMA in real-time. Here the line to parametric rendering becomes blurred, so very similar questions arise as already described in the paragraph on parametric spatial audio.

Perceptual comparison of different binaural renderings in various rooms based on different SMAs, mitigation approaches, and rendering methods (i.e., impulse-response based compared to real-time implementations) would be a highly interesting study. Such a study would bring together current research results and provide a good overview of the application areas of the different methods. Moreover, quality-based experiments on estimated thresholds for the minimum required SH order for different sound field components are of interest. Such experiments would most likely show that the spatial resolution can be further reduced for application-oriented uncritical listening without significant perceptual impairments.

Future research on human voice directivity will focus on estimating the dynamic directivity patterns from fluent speech. This requires adapting recently developed methods for estimating a high-resolution full-spherical voice directivity, as presented in this thesis, for block-based real-time processing. Furthermore, there is still a need for research on how dynamic directivity patterns can be integrated into rendering pipelines. Rendering speech with dynamic voice directivity is generally possible if the spoken sentence is known and the corresponding directivity is prepared. However, for real-time applications in which, for example, a speaker in the form of a digital avatar speaks freely in a VR/AR scene, there are no solutions yet for integrating the dynamic voice directivity corresponding to the spoken sentences. One possible approach would be to capture the speaker's directivity in real-time with a small egocentric microphone array, apply spatial upsampling to it, and then use it for rendering. However, such approaches still require further research.

To estimate how much effort in capturing and reproducing is needed for perceptually plausible rendering of speech in VR/AR scenes, psychoacoustic experiments are needed to determine the required minimum spatial resolution (or, in general, the required accuracy) of voice directivity. Furthermore, psychoacoustic experiments are needed to examine whether a detailed reproduction of individual voice directivity is necessary or whether an approximation (e.g., by an analytical model) leads to perceptually satisfactory or even similarly plausible results. All such experiments are ideally conducted in a 6-DoF environment where participants can walk around a virtual speaker, resulting in clearly audible voice directivity effects.

1.6 References

Ahrens, J. (2019). Auralization of omnidirectional room impulse responses based on the spatial decomposition method and synthetic spatial data. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK* (pp. 146–150). https://doi.org/10.1109/ICASSP.2019.8683661

Ahrens, J., & Andersson, C. (2019). Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre. *J. Acoust. Soc. Am.*, *145*(4), 2783–2794. https://doi.org/10.1121/1.5096164

Algazi, V. R., Avendano, C., & Duda, R. O. (2001). Estimation of a Spherical-Head Model from Anthropometry. *J. Audio Eng. Soc.*, *49*(6), 472–479.

Amengual Garí, S. V., Brimijoin, W. O., Hassager, H. G., & Robinson, P. W. (2019). Flexible binaural resynthesis of room impulse responses for augmented reality research. In *Proc. of the EAA Spatial Audio Signal Processing Symposium, Paris, France* (pp. 161–166). https://doi.org/10.25836/sasp.2019.31

Avni, A., Ahrens, J., Geier, M., Spors, S., Wierstorf, H., & Rafaely, B. (2013). Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution. *J. Acoust. Soc. Am.*, *133*(5), 2711–2721. https://doi.org/10.1121/1.4795780

Bech, S. (1998). Spatial aspects of reproduced sound in small rooms. *J. Acoust. Soc. Am.*, *103*(1), 434–445. https://doi.org/10.1121/1.421098

Ben-Hur, Z., Brinkmann, F., Sheaffer, J., & Weinzierl, S. (2017). Spectral equalization in binaural signals represented by order-truncated spherical harmonics. *J. Acoust. Soc. Am.*, *141*(6), 4087–4096. https://doi.org/10.1121/1.4983652

Ben-Hur, Z., Sheaffer, J., & Rafaely, B. (2018). Joint sampling theory and subjective investigation of plane-wave and spherical harmonics formulations for binaural reproduction. *Appl. Acoust.*, *134*, 138–144. https://doi.org/10.1016/j.apacoust.2018.01.016

Bernschütz, B. (2013). A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100. In *Proc. of the 39th DAGA, Merano, Italy* (pp. 592–595).

Bernschütz, B. (2016). *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording* (Doctoral Dissertation, TU Berlin). http://dx.doi.org/10.14279/depositonce-5082

Bernschütz, B., Giner, A. V., Pörschmann, C., & Arend, J. M. (2014). Binaural Reproduction of Plane Waves With Reduced Modal Order. *Acta Acust. united Ac.*, *100*(5), 972–983. https://doi.org/10.3813/AAA.918777

Blauert, J. (1996). *Spatial Hearing - The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press.

Brinkmann, F., Aspöck, L., Ackermann, D., Lepa, S., Vorländer, M., & Weinzierl, S. (2019). A round robin on room acoustical simulation and auralization. *J. Acoust. Soc. Am.*, *145*(4), 2746–2760. https://doi.org/10.1121/1.5096178

Brinkmann, F., Dinakaran, M., Pelzer, R., Grosche, P., Voss, D., & Weinzierl, S. (2019). A Cross-Evaluated Database of Measured and Simulated HRTFs Including 3D Head Meshes, Anthropometric Features, and Headphone Impulse Responses. *J. Audio Eng. Soc.*, *67*(9), 705–718. https://doi.org/10.17743/jaes.2019.0024

Brinkmann, F., Gamper, H., Raghuvanshi, N., & Tashev, I. (2020). Towards encoding perceptually salient early reflections for parametric spatial audio rendering. In *Proc. of the 148th AES Convention, Vienna, Austria* (pp. 1–11).

Brinkmann, F., Lindau, A., & Weinzierl, S. (2017). On the authenticity of individual dynamic binaural synthesis. *J. Acoust. Soc. Am.*, *142*(4), 1784–1795. https://doi.org/10.1121/1.5005606

Brown, A. D., Stecker, G. C., & Tollin, D. J. (2015). The Precedence Effect In Sound Localization. *JARO*, *16*(1), 1–28. https://doi.org/10.1007/s10162-014-0496-2

Brungart, D. S. (1999). Auditory localization of nearby sources. III. Stimulus effects. *J. Acoust. Soc. Am.*, *106*(6), 3589–3602. https://doi.org/10.1121/1.428212

Brungart, D. S., & Rabinowitz, W. M. (1999). Auditory localization of nearby sources. Head-related transfer functions. *J. Acoust. Soc. Am.*, *106*(3), 1465–1479. https://doi.org/10.1121/1.427180

Coleman, P., Franck, A., Jackson, P. J., Remaggi, L., & Melchior, F. (2017). Object-Based Reverberation for Spatial Audio. *J. Audio Eng. Soc.*, *65*(1/2), 66–77. https://doi.org/10.17743/jaes.2016.0059

Duraiswami, R., Zotkin, D. N., & Gumerov, N. A. (2004). Interpolation and Range Extrapolation of HRTFs. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Montreal, Quebec, Canada* (pp. IV45–IV48). https://doi.org/10.1109/ICASSP.2004.1326759

Engel, I., Henry, C., Amengual Garí, S. V., Robinson, P. W., & Picinali, L. (2021). Perceptual implications of different Ambisonics-based methods for binaural reverberation. *J. Acoust. Soc. Am.*, *149*(2), 895–910. https://doi.org/10.1121/10.0003437

Erraji, A., Stienen, J., & Vorländer, M. (2021). The image edge model. *Acta Acust.*, *5*(17), 1–15. https://doi.org/10.1051/aacus/2021010

Fallahi, M., Hansen, M., Doclo, S., Van De Par, S., Püschel, D., & Blau, M. (2021). Evaluation of head-tracked binaural auralizations of speech signals generated with a virtual artificial head in anechoic and classroom environments. *Acta Acustica*, *5*(30), 1–18. https://doi.org/10.1051/aacus/2021025

Gardner, W. G., & Keith, D. M. (1995). HRTF measurements of a KEMAR. *J. Acoust. Soc. Am.*, *97*(6), 3907–3908. https://doi.org/10.1121/1.412407

Hartmann, W. M. (1983). Localization of sound in rooms. *J. Acoust. Soc. Am.*, *74*(5), 1380–1391. https://doi.org/10.1121/1.390163

Helmholz, H., Andersson, C., & Ahrens, J. (2019). Real-Time Implementation of Binaural Rendering of High-Order Spherical Microphone Array Signals. In *Proc. of the 45th DAGA, Rostock, Germany* (pp. 1462–1465).

Jot, J.-M., Larcher, V., & Pernaux, J.-M. (1999). A Comparative Study of 3-D Audio Encoding and Rendering Techniques. In *Proc. of the 16th International AES Conference on Spatial Sound Reproduction, Rovaniemi, Finland* (pp. 281–300).

Kan, A., Jin, C., & van Schaik, A. (2009). A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function. *J. Acoust. Soc. Am.*, *125*(4), 2233–2242. https://doi.org/10.1121/1.3081395

Kaneko, S., & Duraiswami, R. (2021). Multiple scattering ambisonics: Three-dimensional sound field estimation using interacting spheres. *JASA Express Lett.*, *1*(8), 1–6. https://doi.org/10.1121/10.0005832

Kleiner, M., Dalenbäck, B.-I., & Svensson, U. P. (1993). Auralization - An Overview. *J. Audio Eng. Soc.*, *41*(11), 861–875.

Kolarik, A. J., Moore, B. C. J., Zahorik, P., Cirstea, S., & Pardhan, S. (2016). Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Atten. Percept. Psychophys.*, *78*(2), 373–395. https://doi.org/10.3758/s13414-015-1015-1

Kopčo, N., Doreswamy, K. K., Huang, S., Rossi, S., & Ahveninen, J. (2020). Cortical auditory distance representation based on direct-to-reverberant energy ratio. *NeuroImage*, *208*, 116436. https://doi.org/10.1016/j.neuroimage.2019.116436

Lindau, A. (2014). *Binaural Resynthesis of Acoustical Environments. Technology and Perceptual Evaluation.* (Doctoral Dissertation, TU Berlin). http://dx.doi.org/10.14279/depositonce-4085

Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., & Weinzierl, S. (2014).

A spatial audio quality inventory (SAQI). *Acta Acust. united Ac.*, *100*(5), 984–994. https://doi.org/10.3813/AAA.918778

Lindau, A., & Weinzierl, S. (2012). Assessing the Plausibility of Virtual Acoustic Environments. *Acta Acust. united Ac.*, *98*(5), 804–810. https://doi.org/10.3813/AAA.918562

Litovsky, R. Y., Colburn, H. S., Yost, W. A., & Guzman, S. J. (1999). The precedence effect. *J. Acoust. Soc. Am.*, *106*(4), 1633–1654. https://doi.org/10.1121/1.427914

Madmoni, L., Donley, J., Tourbabin, V., & Rafaely, B. (2020). Beamforming-based Binaural Reproduction by Matching of Binaural Signals. In *Proc. of the AES International Conference on Audio for Virtual and Augmented Reality (AVAR), Redmond, WA, USA* (pp. 1–8).

McCormack, L., & Politis, A. (2019). SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods. In *Proc. of the AES International Conference on Immersive and Interactive Audio, York, UK* (pp. 1–12).

McKeag, A., & McGrath, D. (1996). Sound Field Format to Binaural Decoder with Head Tracking. In *Proc. of the 6th AES Australian Regional Convention, Melbourne, Australia* (pp. 1–9).

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

Møller, H. (1992). Fundamentals of Binaural Technology. *Appl. Acoust.*, *36*(3/4), 171–218. https://doi.org/10.1016/0003-682X(92)90046-U

Møller, H., Sørensen, M. F., Hammershøi, D., & Jensen, C. B. (1995). Head-Related Transfer Functions of Human Subjects. *J. Audio Eng. Soc.*, *43*(5), 300–321.

Møller, H., Sørensen, M. F., Jensen, C. B., & Hammershøi, D. (1996). Binaural technique: do we need individual recordings? *J. Audio Eng. Soc.*, *44*(6), 451–469.

Müller, K., & Zotter, F. (2020). Auralization based on multi-perspective ambisonic room impulse responses. *Acta Acust.*, *4*(25), 1–18. https://doi.org/10.1051/aacus/2020024

Okano, T., Beranek, L. L., & Hidaka, T. (1998). Relations among interaural cross-correlation coefficient (IACCE), lateral fraction (LFE), and apparent source width (ASW) in concert halls. *J. Acoust. Soc. Am.*, *104*(1), 255–265. https://doi.org/10.1121/1.423955

Patricio, E., Rumiński, A., Kuklasiński, A., Januszkiewicz, Ł., & Żernicki, T. (2019). Toward Six Degrees of Freedom Audio Recording and Playback Using Multiple Ambisonics Sound Fields. In *Proc. of the 146th AES Convention, Dublin, Ireland* (pp. 1–9).

Pihlajamäki, T., & Pulkki, V. (2015). Synthesis of Complex Sound Scenes with Transformation of Recorded Spatial Sound in Virtual Reality. *J. Audio Eng. Soc.*, *63*(7/8), 542–551. http://dx.doi.org/10.17743/jaes.2015.0059

Pollow, M., Nguyen, K.-V., Warusfel, O., Carpentier, T., Müller-Trapet, M., Vorländer, M., & Noisternig, M. (2012). Calculation of Head-Related Transfer Functions for Arbitrary Field Points Using Spherical Harmonics Decomposition. *Acta Acust. united Ac.*, *98*(1), 72–82. https://doi.org/10.3813/AAA.918493

Pulkki, V. (2007). Spatial Sound Reproduction with Directional Audio Coding. *J. Audio Eng. Soc.*, *55*(6), 503–516.

Pulkki, V., Delikaris-Manias, S., & Politis, A. (2018). *Parametric Time–Frequency Domain Spatial Audio* (1st ed.). Hoboken, NJ, USA: John Wiley & Sons Ltd. https://doi.org/10.1002/9781119252634

Rafaely, B. (2015). *Fundamentals of Spherical Array Processing*. Berlin, Germany: Springer. https://doi.org/10.1007/978-3-662-45664-4

Richter, J.-G. (2019). *Fast Measurement of Individual Head-Related Transfer Functions* (Doctoral Dissertation, RWTH Aachen). https://doi.org/10.30819/4906

Savioja, L., & Svensson, U. P. (2015). Overview of geometrical room acoustic modeling techniques. *J. Acoust. Soc. Am.*, *138*(2), 708–730. https://doi.org/10.1121/1.4926438

Schissler, C., Stirling, P., & Mehra, R. (2017). Efficient Construction of the Spatial Room Impulse Response. In *Proc. of the IEEE Virtual Reality (VR)* (pp. 122–130). https://doi.org/10.1109/VR.2017.7892239

Schröder, D. (2011). *Physically Based Real-Time Auralization of Interactive Virtual Environments* (Doctoral Dissertation). RWTH Aachen.

Spagnol, S., Tavazzi, E., & Avanzini, F. (2017). Distance rendering and perception of nearby virtual sound sources with a near-field filter model. *Appl. Acoust.*, *115*, 61–73. https://doi.org/10.1016/j.apacoust.2016.08.015

Stade, P., Arend, J. M., & Pörschmann, C. (2017). Perceptual Evaluation of Synthetic Early Binaural Room Impulse Responses Based on a Parametric Model. In *Proc. of the 142nd AES Convention, Berlin, Germany* (pp. 1–10).

Stade, P., Bernschütz, B., & Rühl, M. (2012). A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios. In *Proc. of the 27th Tonmeistertagung - VDT International Convention, Cologne, Germany* (pp. 1–17).

Stern, R. M., Wang, D., & Brown, G. J. (2006). Binaural Sound Localization. In D. Wang & G. J. Brown (Eds.), *Computational Auditory Scene Analysis: Principles, Algorithms and Applications* (pp. 147–185). New York: Wiley-IEEE Press.

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

Thurlow, W. R., & Runge, P. S. (1967). Effect of Induced Head Movements on Localization of Direction of Sounds. *J. Acoust. Soc. Am.*, *42*(2), 480–488. https://doi.org/10.1121/1.1910604

Toole, F. E. (2008). *Sound Reproduction - Loudspeakers and Rooms* (1st ed.). Burlington, MA: Focal Press. https://doi.org/10.4324/9780080888019

Vorländer, M. (2008). *Auralization - Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality* (1st ed.). Berlin Heidelberg: Springer. https://doi.org/10.1007/978-3-540-48830-9

Werner, S., Klein, F., & Götz, G. (2019). Investigation on spatial auditory perception using non-uniform spatial distribution of binaural room impulse responses. In *Proc. of the International Conference on Spatial Audio (ICSA)* (pp. 137–144). https://doi.org/10.22032/dbt.39967

Werner, S., Neidhardt, A., Klein, F., & Brandenburg, K. (2018). Comparison of Different Methods to Create an Interactive Augmented Auditory Reality Scenario Using Sparse Binaural Room Impulse Response Measurements. In *Proc. of the 44th DAGA, Munich, Germany* (pp. 1–4).

Wightman, F. L., & Kistler, D. J. (1989). Headphone simulation of free-field listening. I: Stimulus synthesis. *J. Acoust. Soc. Am.*, *85*(2), 858–867. https://doi.org/10.1121/1.397557

Wightman, F. L., & Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *J. Acoust. Soc. Am.*, *91*(3), 1648–1661. https://doi.org/10.1121/1.402445

Xie, B., Zhong, X., Yu, G., Guan, S., Rao, D., Liang, Z., & Zhang, C. (2013). Report on Research Projects on Head-Related Transfer Functions and Virtual Auditory Displays in China. *J. Audio Eng. Soc.*, *61*(5), 314–326.

Zahorik, P. (2021). Spatial Hearing in Rooms and Effects of Reverberation. In R. Y. Litovsky, M. J. Goupell, R. R. Fay, & A. N. Popper (Eds.), *Binaural Hearing* (pp. 243–280). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-57100-9

Zahorik, P., Brungart, D. S., & Bronkhorst, A. W. (2005). Auditory Distance Perception in Humans: A Summary of Past and Present Research. *Acta Acust. united Ac.*, *91*(3), 409–420.

Zaunschirm, M., Frank, M., & Zotter, F. (2018). BRIR synthesis using first-order microphone arrays. In *Proc. of the 144th AES Convention, Milan, Italy* (pp. 1–10).

Zaunschirm, M., Schoerkhuber, C., & Hoeldrich, R. (2018). Binaural rendering of Ambisonic signals

by HRIR time alignment and a diffuseness constraint. *J. Acoust. Soc. Am.*, *143*(6), 3616–3627. https://doi.org/10.1121/1.5040489

Ziegelwanger, H., Majdak, P., & Kreuzer, W. (2015). Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization. *J. Acoust. Soc. Am.*, *138*(1), 208–222. https://doi.org/10.1121/1.4922518

Zotter, F., & Frank, M. (2012). All-round ambisonic panning and decoding. *J. Audio Eng. Soc.*, *60*(10), 807–820.

Zotter, F., & Frank, M. (2019). *Ambisonics: A Practical 3D Audio Theory for Recording*. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-17207-7

2 Head-Related Transfer Function Processing and Interpolation

2.1 Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions

Arend, J. M., Brinkmann, F., & Pörschmann, C. (2021). *J. Audio Eng. Soc.*, *69*(1/2), 104–117. https://doi.org/10.17743/jaes.2020.0070

OPEN ACCESS Freely available online

J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions"
J. Audio Eng. Soc., vol. 69, no. 1/2, pp. 104–117, (2021 January/February).
DOI: https://doi.org/10.17743/jaes.2020.0070

Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions

JOHANNES M. AREND,[1,2] *AES Student Member*, **FABIAN BRINKMANN,**[2] *AES Associate Member* **AND**
(johannes.arend@th-koeln.de) (fabian.brinkmann@tu-berlin.de)

CHRISTOPH PÖRSCHMANN,[1] *AES Associate Member*
(christoph.poerschmann@th-koeln.de)

[1]*TH Köln – University of Applied Sciences, Cologne, Germany*
[2]*Technical University of Berlin, Berlin, Germany*

High-quality spatial audio reproduction over headphones requires head-related transfer functions (HRTFs) with high spatial resolution. However, acquiring datasets with a large number of (individual) HRTFs is not always possible, and using large datasets can be problematic for real-time applications with limited resources. Consequently, interpolation methods for sparsely sampled HRTFs are of great interest, with spherical harmonics (SH) interpolation becoming increasingly popular. However, the SH representation of sparse HRTFs suffers from spatial aliasing and order truncation errors. To mitigate this, preprocessing methods have been introduced that time-align the sparse HRTFs before SH interpolation. This reduces the effective SH order and thus the number of HRTFs required for SH interpolation. In this paper, we present a physical evaluation of four state-of-the-art preprocessing methods, which showed very similar performance of the methods with notable differences only at low SH orders and contralateral HRTFs. We also performed a listening experiment with one selected method to determine the minimum required SH order required for perceptually transparent interpolation. For the selected method, a sparse HRTF set of order $N \approx 7$ is sufficient for interpolating a frontal source presenting speech or percussion. Higher orders are, however, required for a lateral source and noise.

0 INTRODUCTION

Head-related transfer functions (HRTFs) are one key component for headphone-based spatial audio rendering, as often used in virtual reality (VR) or augmented reality (AR) applications [1, 2]. HRTFs describe the sound incidence from a source to the left and right ear and the associated directional filtering of incoming sound by the pinna, head, and torso. As such, HRTFs include binaural cues (i.e., interaural level differences (ILDs) and interaural time differences (ITDs) primarily used for sound source localization in the horizontal plane) as well as monaural spectral cues primarily used for sound source localization in the median plane [3].

For high-quality spatial audio over headphones, HRTFs with high spatial resolution are essential. Usually, such data are measured on dense spherical sampling grids, which can be achieved by sequential measurements to obtain dummy head HRTFs [4–6], but require procedures and equipment optimized for speed if measuring human subjects [7–9]. For this purpose, measurement systems consisting of (semi)circular loudspeaker arcs are used with signal acquisition techniques that allow for a continuous rotation of the subject or arc. Given this, it is of great interest to measure fewer HRTFs on a sparse spatial sampling grid and generate dense HRTF sets by means of interpolation (also referred to as spatial upsampling). This would decrease the cost and complexity of HRTF measurement systems and allow for faster rotations depending on the acquisition method. Furthermore, interpolation of sparse HRTF sets may reduce the memory and computational load for real-time applications with limited resources (e.g., mobile applications).

Currently very popular is the description and interpolation of HRTFs in the spatially continuous spherical harmonics (SH) domain (see Sec. 1). However, the required number of spatial samples (i.e., measurement directions) increases with frequency, and an SH order (also called spatial order) of $N_{\mathrm{max}} \approx 40$ is needed for a physically correct interpolation up to 20 kHz, resulting in at least $(N + 1)^2 = 1{,}681$ measurement directions [10]. Obviously, sparse HRTF sets

do not meet this requirement, and their SH representation thus suffers from so-called sparsity errors, which is a combination of spatial aliasing and order truncation errors [11]. Because sparse sampling grids only allow SH processing up to $N_{\text{sparse}} < N_{\text{max}}$, energy above N_{sparse} is irreversibly aliased to lower orders, causing spatial ambiguities that result in a high-shelf-like energy increase in SH interpolated HRTFs [12, 13, 11]. The predominant truncation error leads to reduced spatial detail showing up as a severe high frequency roll-off [14, 15, 11], caused by discarding energy above N_{sparse}. In combination, these effects also result in ILD errors and degraded loudness stability in dynamic scenes [16, 11].

To enable accurate SH interpolation of sparse HRTF sets, several preprocessing techniques have been introduced. In the present study, we focus on methods that align the head-related impulse responses (HRIR, time-domain equivalent of the HRTF) in the time [17, 18] or frequency domain [15, 19, 16, 20] prior to the SH interpolation and reverse the alignment afterwards. Since most higher-order HRTF energy stems from rapid spatial phase changes, aligning the HRIRs and thus also the phase components significantly decreases the high-order energy and related sparsity errors [18, 21, 20]. Because the phase changes are caused by the distance of the ears to the coordinate origin—the center of the head, in this case—the alignment can also be interpreted as centering the ears in the origin. For a more comprehensive overview of preprocessing methods, please refer to [19] and Chapter 4.11 of [22].

The studies on preprocessing introduced in the previous paragraph all showed that time-alignment decreases spectral and temporal errors and thus increases the quality of interpolated HRTFs, especially for low-order SH interpolation. However, listening experiments assessing the perceptual performance of SH-based HRTF interpolation either with or without preprocessing are rare. The only study directly related to the topic was presented by Pike and Tew (see Chapter A.8 of [18]). They conducted a Multi-Stimulus Test with Hidden Reference and Anchor (MUSHRA), comparing perceivable differences between a measured reference and SH interpolated HRTFs with and without subsample precise onset-based time-alignment. While interpolated HRTFs were indistinguishable from the reference at $N = 35$ in both cases, at $N = 5$, the time-alignment significantly reduced perceptual differences at least for frontal source positions, whereas for a lateral source position perceptual differences were still clear. Besides that, there are a few studies on the impact of low-order SH representation of HRTFs on localization accuracy [23], perceived loudness stability [11], or speech intelligibility in noise [24].

To the best of our knowledge, a systematic comparison of the different alignment approaches and listening experiments to find the minimum order N that is required for a perceptually transparent SH interpolation is missing so far. Because the methods differ in their computational complexity, a detailed comparison might help to choose the method that is most appropriate for a specific application, whereas the minimum required SH order is of importance for high-quality applications and can provide a starting point for further perceptual studies for applications that allow for a certain quality degradation. To close this gap, we present a physical evaluation of all suggested methods showing that they perform comparably. In addition, we conducted an adaptive forced choice listening experiment with one selected alignment approach to examine the minimum SH order required for interpolated HRTFs to be indistinguishable from a measured reference.

The remainder is structured as follows. Sec. 1 briefly reviews the fundamentals of HRTF representation and interpolation in the SH domain, and Sec. 2 describes the different preprocessing methods in detail. Sec. 3 provides a physical evaluation of the discussed methods by means of spectral and temporal error measures. Sec. 4 describes the listening experiment and results, followed by a discussion and conclusion in Sec. 5 and Sec. 6.

1 SPHERICAL HARMONICS REPRESENTATION OF HRTFS

The HRTF $H^{\text{l, r}}(\omega, \Omega)$ for the left and right ear can be represented in the SH domain by a set of SH coefficients $h_{nm}^{\text{l,r}}(\omega)$, which can be obtained by the spherical Fourier transform (SFT) (see Chapter 1 of [25]) (indices for the left and right ear are omitted in the following whenever the processing is identical for both ears):

$$h_{nm}(\omega) = \int_0^{2\pi} \int_0^{\pi} H(\omega, \Omega) Y_n^m(\Omega)^* \cos\theta \mathrm{d}\theta \mathrm{d}\phi. \tag{1}$$

The angular frequency is given by $\omega = 2\pi f$, with f being the temporal frequency. The direction $\Omega = (\phi, \theta)$ is defined by the azimuth $\phi = [0^\circ, 360^\circ]$ and the elevation $\theta = [-90^\circ, 90^\circ]$, whereby ϕ is measured counterclockwise in the xy-plane, starting at positive x, and θ is 90° at positive z. The notation $(\cdot)^*$ denotes the complex conjugate and Y_n^m the complex SH basis functions of order n and degree m defined as

$$Y_n^m(\theta, \phi) = \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\sin\theta) \mathrm{e}^{\mathrm{i}m\phi}, \tag{2}$$

with the associated Legendre functions P_n^m and the imaginary unit $\mathrm{i} = \sqrt{-1}$.

In practice, the HRTF is sampled at a finite number of directions, and therefore, the integral in Eq. (1) must be discretized to Q sampling points corresponding to the measurement directions Ω_q. The respective discrete SFT is defined as

$$h_{nm}(\omega) = \sum_{q=1}^{Q} \alpha_q \, H(\omega, \Omega_q) Y_n^m(\Omega_q)^*, \tag{3}$$

where the quadrature weights α_q compensate for an uneven distribution of the sampling points (Chapter 4 of [26]). Alternatively, the discrete SFT can also be formulated in matrix form and then calculated by an inversion of the respective SH transformation matrix (Chapter 3 of [25]), but for the present work, the discrete SFT was always calculated using the closed-form expression according to Eq. (3).

Due to the analytical and spatially continuous basis functions, the SH representation allows for interpolation, that is, HRTFs $\widehat{H}(\omega, \Omega_t)$ for any direction Ω_t can be reconstructed by the discrete inverse spherical Fourier transform (ISFT):

$$\widehat{H}(\omega, \Omega_t) = \sum_{n=0}^{N} \sum_{m=-n}^{n} h_{nm}(\omega) Y_n^m(\Omega_t)\,. \tag{4}$$

However, the discrete sampling in Eq. (3) directly limits the maximum resolvable SH order N,

$$N \leq \lfloor \sqrt{Q/\lambda} - 1 \rfloor, \tag{5}$$

with the efficiency factor $\lambda \geq 1$ that depends on the sampling scheme and the floor operator $\lfloor \cdot \rfloor$. Thus, sparsity errors occur if $N < N_{\max} \approx 40$. As mentioned in the introduction, these errors manifest in spatial ambiguities, reduced spatial resolution, and spectral and temporal distortions in the interpolated HRTFs. SH interpolation of the complex HRTF spectra according to Eqs. (3) and (4) will be referred to as unprocessed (UP) interpolation in the following (i.e., SH interpolation without time-alignment).

2 TIME-ALIGNED SPHERICAL HARMONICS INTERPOLATION

This section introduces the investigated methods in depth. Although the algorithms differ in detail, the underlying idea is the same. All algorithms aim to lower the SH order that is required for high-quality SH interpolation by minimizing the phase changes across space during preprocessing. This is always done separately for the left and right ear and is pursued by aligning the impulse responses by means of time or frequency domain processing. In all cases, this is achieved by a spectral multiplication or division of the HRTF with an alignment function. After the alignment, all algorithms perform the discrete SFT and ISFT according to Eqs. (3) and (4) using the complex HRTF spectra. However, the time of arrival (TOA) (i.e., the time where the onsets occur in the HRIRs) is lost during the alignment. Therefore, it has to be reconstructed after the interpolation, which requires a spatially continuous TOA model in postprocessing. To foster reproducible research, example implementations of the methods under investigation are published as part of the SUpDEq Toolbox for MATLAB[1].

2.1 Onset-Based Time-Alignment

Sample accurate onset-based time-alignment (OBTA) was first proposed by Evans et al. [17] and was refined to subsample accuracy by Pike and Tew [18] as well as by Brinkmann and Weinzierl [19]. In preprocessing, the TOAs of the HRIRs are first detected by threshold-based onset detection and then removed using fractional delays. The time-aligned, complex HRTF spectra and the extracted TOAs are then interpolated separately to any desired (dense) sampling grid using Eqs. (3) and (4). Afterwards, the TOA is reconstructed in postprocessing using fractional delays once again.

We implemented the method as described by Brinkmann and Weinzierl [19] using onset detection with a threshold of −20 dB in relation to the maximum values of the 10 times upsampled and low-pass–filtered HRIRs (8th order Butterworth, $f_c = 3$ kHz, see [27]). The TOAs were removed and inserted in frequency domain using fractional delay filters and circular convolution, which has the advantage that the length of the HRIRs is not changed during the processing. The fractional delays were designed in the time domain using Kaiser windowed sinc filters of order 70 with a side lobe attenuation of 60 dB [28]. The filters exhibit negligible magnitude distortions <0.1 dB and group delay distortions <0.1 samples below 20 kHz.

2.2 Frequency-Dependent Time-Alignment

Zaunschirm et al. [15] presented a frequency-dependent time-alignment (FDTA) as HRTF preprocessing for binaural Ambisonics renderering. FDTA removes the high frequency TOA and thus also the ITD above 1.5 kHz and maintains it at low frequencies. Because the ITDs become less relevant as frequency increases [29], the authors proposed not to resynthesize the high-frequency ITDs for binaural reproduction of the Ambisonics signal. However, the alignment can easily be reversed to reconstruct HRTFs after SH interpolation.

In contrast to the onset-based time-alignment, FDTA does not aim to completely remove the TOAs. Instead, TOA differences between HRIRs are removed and a constant TOA remains. We refer to this as relative TOA alignment in the following.

The relative TOAs $\tau_q^{\mathrm{l,r}}$ are estimated from the time difference by which a plane wave from direction Ω_q arrives at the center of the head and the position of the ear

$$\tau_q^{\mathrm{r}} = \cos\theta_q \sin\phi_q r_0 c^{-1}, \qquad \tau_q^{\mathrm{l}} = -\tau_q^{\mathrm{r}}, \tag{6}$$

with $c = 343$ m/s the speed of sound, r_0 the head radius, and q a spatial sampling point of the HRTF. This inherently assumes that the ears are located at $\phi_{\mathrm{e}} = [90^\circ, 270^\circ]$ and $\theta_{\mathrm{e}} = [0^\circ, 0^\circ]$ and neglects diffraction around the head that might affect the actual TOA. The estimated relative TOAs are the basis for designing an all-pass filter $A_q^{l,r}(\omega)$ for each sampling point q, which is applied by multiplication in the frequency domain to achieve the relative TOA alignment. The filter is defined as

$$A_q^{\mathrm{l,r}}(\omega) = \begin{cases} 1 & \text{for } \omega < \omega_c \\ \mathrm{e}^{-\mathrm{i}(\omega-\omega_c)\tau_q^{\mathrm{l,r}}} & \text{for } \omega \geq \omega_c, \end{cases} \tag{7}$$

where $\omega_c = 2\pi f_c$ with the cut-on frequency $f_c = 1.5$ kHz. Thus, the filter exhibits a group delay of 0 below f_c and $\tau_q^{\mathrm{l,r}}$ above.

After SH interpolation of the time-aligned HRTFs to T desired directions Ω_t, the original ITDs can be reconstructed by reversing the alignment. Thus, all-pass filters for each direction t are calculated according to Eqs. (6) and (7) and applied by division in the frequency domain.

[1] Available: https://github.com/AudioGroupCologne/SUpDEq.

2.3 Spatial Upsampling by Directional Equalization

With Spatial Upsampling by Directional Equalization (SUpDEq), we recently presented a method using a rigid sphere as a simplified head model as the basis for the alignment [16, 30–33]. This has the advantage that scattering effects around the head are approximated. As with FDTA, SUpDEq aims at a relative TOA alignment.

In preprocessing, the sparse HRTF set $H(\omega, \Omega_q)$ with Q measurement directions is equalized by spectral division with rigid sphere transfer functions $H_\mathrm{R}(\omega, \Omega_q)$ described as

$$H_\mathrm{R}(\omega, \Omega_q) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n(kr_0) Y_n^m(\Omega_\mathrm{e}) Y_n^m(\Omega_q)^*, \qquad (8)$$

with Ω_e the left and right ear position. The scattering around the rigid sphere is accounted for by

$$d_n(kr_0) = 4\pi \mathrm{i}^n \left[j_n(kr_0) - \frac{j_n'(kr_0)}{h_n^{(2)'}(kr_0)} h_n^{(2)}(kr_0) \right], \qquad (9)$$

with j_n the spherical Bessel function of the first kind, $h_n^{(2)}$ the spherical Hankel function of the second kind, and j_n' and $h_n^{(2)'}$ their derivatives[2]. The rigid sphere transfer functions are calculated at a high spatial order $N \geq 40$ to avoid sparsity errors.

Because the TOA is contained in the spherical head model, the spectral division of the HRTF by H_R automatically yields the time-alignment and additionally aims at equalizing parts of the magnitude response. H_R may be considered as a simplified HRTF set comprising only basic temporal and spectral features. From an information theory point of view, the result of the equalization can thus be understood as the prediction error between the actual HRTFs and the spherical head model, which has a lower SH order than the original HRTF set.

The equalized HRTFs are interpolated in the SH domain to T desired directions Ω_t using Eqs. (3) and (4). In postprocessing, the interpolated HRTFs are de-equalized by spectral multiplication with rigid sphere transfer functions for the interpolated directions Ω_t to recover previously discarded temporal and spectral components of the HRTF. To maintain valid HRTF data, the equalization and de-equalization were applied in the present study only above the spatial aliasing frequency $f_A = N_s c / 2\pi r_0$, where N_s is the SH order of the sparse sampling grid [35]. This was done by setting $H_\mathrm{R}(\omega, \Omega_q) = 1$ for $0 \leq \omega \leq 2\pi f_A 2^{-1/3}$, where $2^{-1/3}$ represents a third-octave safety margin.

2.4 Phase-Correction

Ben-Hur et al. [20] presented a pre- and postprocessing technique called phase-correction (PC) that is conceptually similar to SUpDEq. In preprocessing, the HRTF set $H(\omega, \Omega_q)$ measured for Q sampling points is equalized by spectral division with open sphere transfer functions $H_\mathrm{O}(\omega, \Omega_q)$ for the corresponding directions Ω_q, given by

$$H_\mathrm{O}(\omega, \Omega_q) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n(kr_0) Y_n^m(\Omega_\mathrm{e}) Y_n^m(\Omega_q)^*, \qquad (10)$$

with

$$d_n(kr_0) = 4\pi \mathrm{i}^n j_n(kr_0) \qquad (11)$$

Compared to Eq. (9), the open sphere transfer function in Eq. (11) does not contain a scattering term and thus results in a frequency-independent time-alignment not accounting for magnitude effects. Therefore, the equalization can also be described as a frequency domain multiplication of the HRTF set $H(\omega, \Omega_q)$ with a phase-correction term (all-pass), which the authors defined as

$$C_q^{\mathrm{l,r}}(\omega) = \mathrm{e}^{-\mathrm{i}kr_0 \cos \Theta_q^{\mathrm{l,r}}}, \qquad (12)$$

where $\Theta_q^{\mathrm{l,r}}$ is the angle between the measured direction Ω_q and the left and right ear position Ω_e and $\cos \Theta_q^{\mathrm{l,r}} = \cos \theta_q \cos \theta_\mathrm{e} + \cos(\phi_q - \phi_\mathrm{e}) \sin \theta_q \sin \theta_\mathrm{e}$.

Applying the phase-correction to the HRTFs in preprocessing results in a time-aligned HRTF set with lower SH order (also referred to as ear-alignment in [20]). After SH interpolation of the phase-corrected HRTFs to T desired directions Ω_t using Eqs. (3) and (4), HRTFs can be reconstructed by applying the inverse phase correction. Thus, phase-correction terms for each direction t are calculated according to Eq. (12) and applied to the interpolated HRTFs by spectral division in the frequency domain.

3 PHYSICAL EVALUATION

The physical evaluation focuses on two aspects: The alignment and restoration of the TOAs as the main methodological difference between the algorithms, and the spectral distortion identified in a previous study as the most problematic artifact [19]. Both aspects are also highly relevant from a perceptual point of view: the TOA is directly related to the ITD, which is the main cue for left/right localization [29], while the perceived coloration and up/down localization errors are attributable to spectral distortions [36].

3.1 HRTFs

HRTFs from a Neumann KU100 measured on a Lebedev grid with 2,702 sampling points [5] were used as the reference allowing for SH interpolation of order $N = 44$ without any sparsity errors. Sparse HRTF sets were then generated by spatially subsampling the reference in the SH domain to Lebedev grids of order $1 \leq N \leq 15$ according to Eqs. (3) and (4). In the last step, the sparse sets were subjected to the processing methods introduced above. Throughout this study, a head radius of $r_0 = 9.19$ cm was used, calculated according to Algazi et al. [37], and the left and right ear position Ω_e required for SUpDEq and PC was defined with $\phi_\mathrm{e} = [90^\circ, 270^\circ]$ and $\theta_\mathrm{e} = [0^\circ, 0^\circ]$.

[2]Please note the dependency of Eq. (9) on the Fourier transform kernel [34, Table I]. We used $p(\omega) = \int_{-\infty}^{\infty} p(t) e^{-\mathrm{i}\omega t} \mathrm{d}t$ as the Fourier transform of the pressure signal $p(t)$.

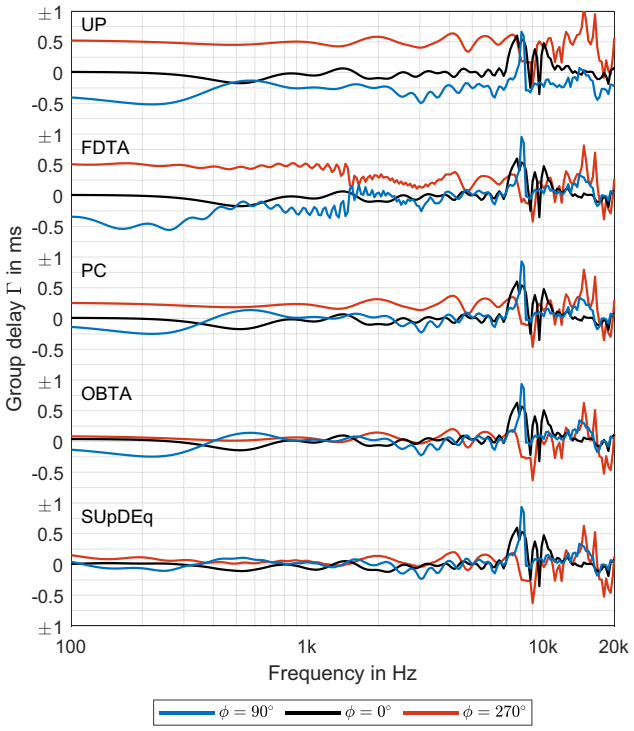


Fig. 1. Group delay of time-aligned HRTFs compared to unprocessed (UP) HRTFs for the left ear and three selected source positions in the horizontal plane ($\theta = 0°$).

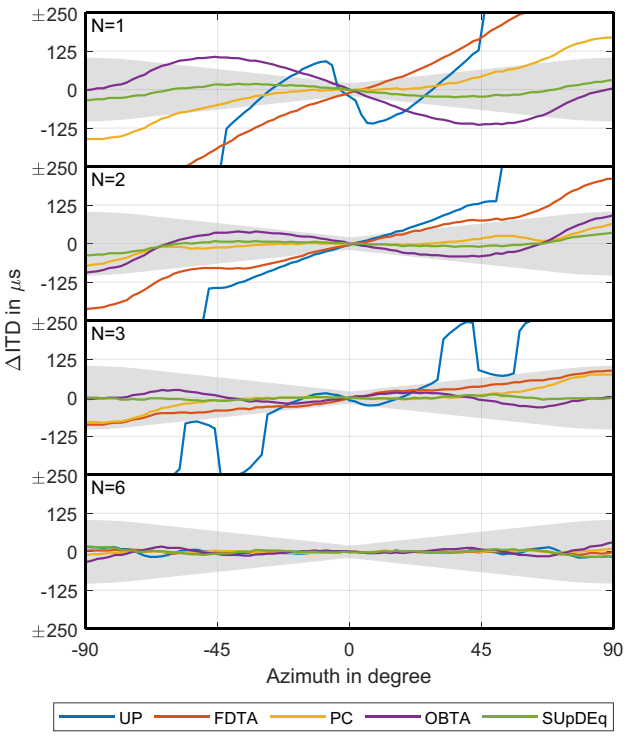


Fig. 2. Difference in ITD relative to the reference for the frontal region of the horizontal plane and selected SH orders N. The shaded area denotes the JND as a function of the reference ITD.

3.2 TOA Alignment

Perfectly aligned HRTFs would show a constant group delay independent of frequency and source position. To assess the performance of the alignment approaches, we calculated the HRTF group delay

$$\Gamma(\omega, \Omega_q) = -\frac{\mathrm{d}\angle H(\omega, \Omega_q)}{\mathrm{d}\omega} \tag{13}$$

for all Q measurement directions, where $\angle H(\cdot)$ is the unwrapped phase response. Because most methods perform a relative alignment, the group delay was centered around 0 ms by subtracting the overall mean separately for each method. Fig. 1 shows group delays of three selected HRTFs in the horizontal plane before the alignment (UP) and after the respective alignment. The unprocessed HRTFs show group delay differences of approximately 1 ms at low frequencies and 0.75 ms at high frequencies. Narrow group delay peaks occur for frequencies above 7 kHz caused by rapid phase changes due to HRTF notches (see also Fig. 3).

As expected, FDTA maintains the group delays below 1.5 kHz and aligns the data for higher frequencies. However, the preprocessing leads to ripples around 1.5 kHz, probably caused by the discontinuity in the alignment function defined in Eq. (7) and the finite HRIR length (Gibbs phenomenon, Chapter 7.5 of [38]). A smooth transition between the two states of the alignment function or windowing the time signal might reduce these ripples. Furthermore, FDTA fails in aligning the contralateral HRTF ($\phi = 270°$) between 1.5 and 8 kHz.

Results for PC are visually very similar to FDTA above 1.5 kHz—apart from the FDTA ripples—which is not surprising, as both methods estimate the TOA based on an open sphere geometry. Below 1.5 kHz, group delay differences of approximately 0.50 ms remain uncompensated because the open sphere TOA alignment does not account for low-frequency phase effects that occur due to the scattering around the head (a visualization of this effect is given in Fig. 2 of [39]).

Low-frequency group delay differences of approximately 0.25 ms remain for OBTA, which is about half of the differences observed for PC. The remaining differences below 1 kHz are mainly caused by the ipsilateral HRTF ($\phi = 90°$), which shows the strongest fluctuations in this range already for UP. Above 1 kHz, OBTA outperforms FDTA and PC due to a better alignment of the contralateral HRTF.

SUpDEq processing yields the smallest group delay deviations across source positions, reducing low-frequency group delay differences in Fig. 1 to about 0.125 ms. This improvement is clearly related to considering the scattering around the sphere in the alignment process. For frequencies above 1 kHz, SUpDEq and OBTA perform comparably well.

An additional analysis of the group delay standard deviation across all source positions, presented in the supplementary material (Fig. S1 of [40]), confirmed the trends observed for the three selected positions. SUpDEq outperforms all remaining methods up to approximately 1.5 kHz. Above 1.5 kHz, all alignment approaches produce comparable standard deviations that remain below that of the unprocessed HRTFs up to 20 kHz.

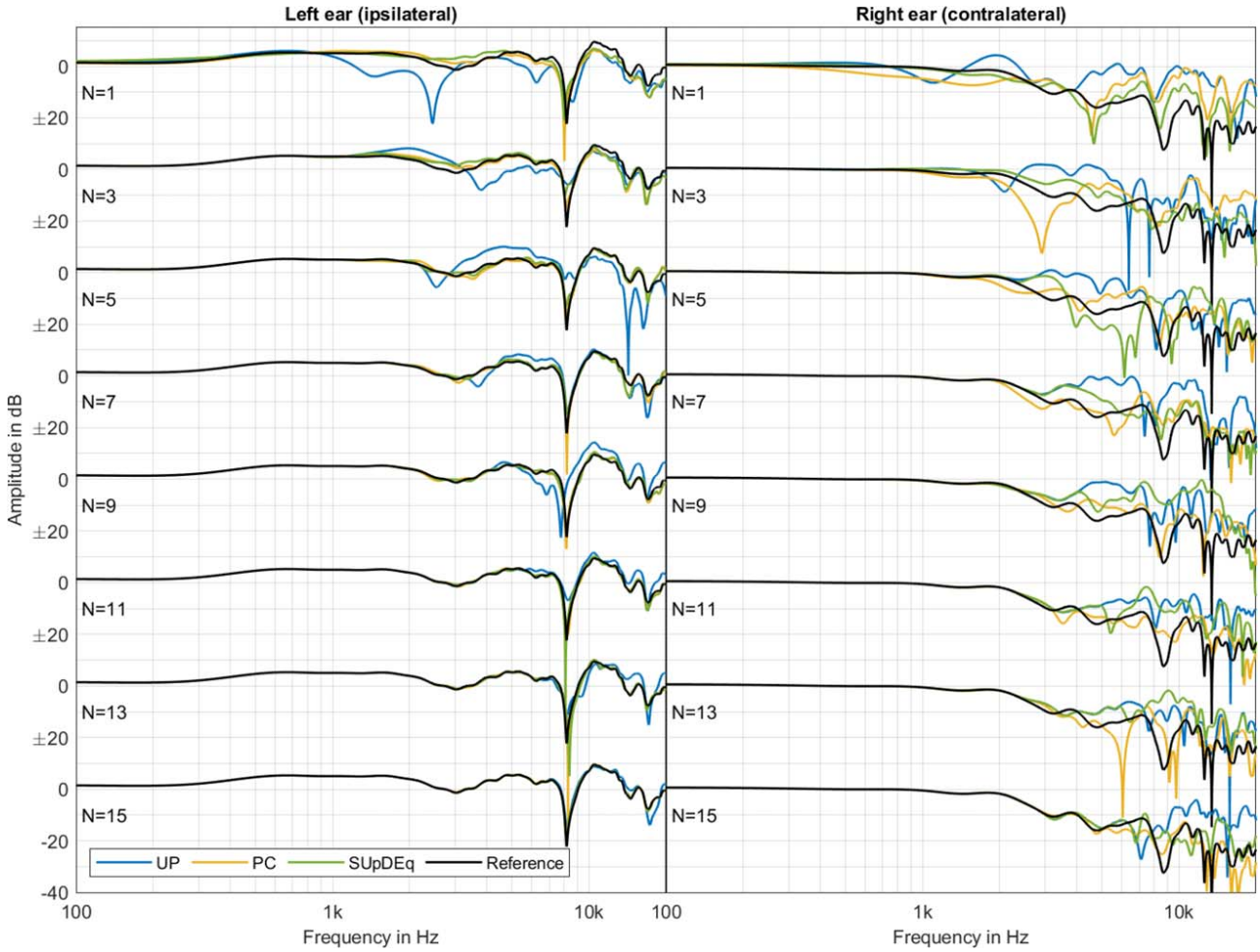


Fig. 3. Reference and interpolated HRTFs for a source at $\Omega = (90°, 0°)$, selected SH orders N and interpolation methods.

3.3 TOA Restoration

To assess the TOA restoration, the horizontal plane ITD was calculated from the difference between the left and right ear TOAs for HRTFs processed with all methods introduced above and SH orders $1 \leq N \leq 15$. The TOAs were estimated from the 10 times upsampled and low-passed HRIRs (8th order Butterworth, $f_c = 3kHz$, see [27]). A threshold of -10 dB was used for TOA detection in all cases. Using a threshold of -30 dB or -20 dB, as recommended by Andreopoulou and Katz [27], would lead to erroneous detections due to preringing in HRIRs processed at low SH orders [16]. Fig. 2 shows the results for selected SH orders by means of differences to the reference ITD for the frontal region of the horizontal plane (results for the rear were almost identical). The gray area denotes the broadband just noticeable difference (JND) as a function of the reference ITD [41]. The JND was linearly interpolated/extrapolated between 20 μs at $\text{ITD}_{\text{ref}} = 0$ μs and 100 μs at $\text{ITD}_{\text{ref}} = 700$ μs.

For first-order SH interpolation, only SUpDEq manages to keep the ITD errors below the JND, most likely due to the consideration of low-frequency scattering effects described above. While errors only slightly exceed the JND for OBTA and PC in this case, large errors are observed for UP and FDTA. For OBTA and PC, the errors fall below the JND at SH order two, while FDTA requires order three. Thus, starting at SH order three, all alignment methods perform comparably well and yield correct ITDs. However, UP still shows large errors at order three and sudden jumps that are caused by preringing in the HRIRs [16], which can, for example, be reduced by SH tapering [42]. At an SH order of six, the errors finally fall below the JND for all methods.

3.4 Spectral Distortion

To get a first impression of the spectral distortion, Fig. 3 shows HRTFs for two source position at selected SH orders and for selected methods (the supplementary material contains figures for all methods [40, Figs. S2–S4]). For the ipsilateral ear, the errors quickly decrease with increasing SH order and HRTFs are already quite similar to the reference at $N = 3$ for PC and SUpDEq. Results for UP are clearly worse, where high-frequency differences remain up to $N = 15$. Errors are generally larger for the contralateral ear and appear to be less predictable in this case. For example, SUpDEq shows a relatively small error at $N = 3$, where it outperforms UP and PC. At $N = 13$, however, the error for SUpDEq is larger than at $N = 3$ and SUpDEq is outperformed by UP and PC in this case.

For a more systematic analysis, the spectral distortion was calculated as the absolute energetic difference between

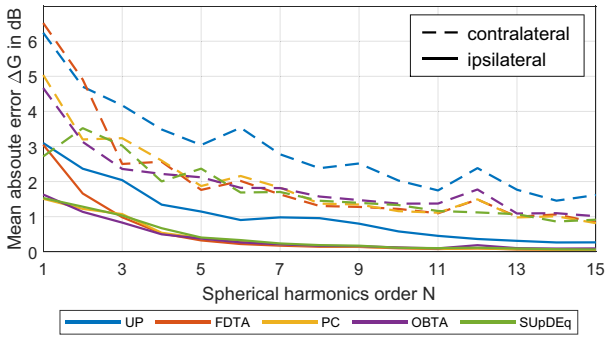


Fig. 4. Left ear energetic error ΔG vs. SH order averaged across frequency and source position in the ipsilateral and contralateral region.

interpolated HRTFs H_i and the reference H_r in 40 auditory filters as implemented in the Auditory Toolbox [43]

$$\Delta G(f_c, \Omega) = \left| 10 \log_{10} \frac{\sum_k C(f_k, f_c)|H_i(f_k, \Omega)|^2}{\sum_k C(f_k, f_c)|H_r(f_k, \Omega)|^2} \right|, \tag{14}$$

with $C(\cdot)$ the auditory filter and f_c the center frequency of the auditory filters and $50\,Hz \leq f_k, f_c \leq 20\,kHz$. The measure $\Delta G(f_c, \Omega)$ was calculated for 900 source positions on a Fliege sampling grid obtained with the SOFiA Toolbox [44]. In the following, averaged errors are denoted by omitting the corresponding symbol, i.e., $\Delta G(f_c)$ gives the error averaged across source position, $\Delta G(\Omega)$ is the frequency average, and ΔG is averaged across source positions and frequencies. Averaging across source positions was done using the quadrature weights α of the Fliege sampling grid.

Fig. 4 shows the left ear errors for ipsilateral and contralateral source regions for all methods and SH orders up to $N = 15$. The errors were obtained by averaging across source positions within 25° great circle distance from $\Omega_{\text{ipsi}} = (90°, 0°)$ and $\Omega_{\text{contra}} = (270°, 0°)$. The supplementary material contains another figure showing errors averaged across all source positions [40, Fig. S5]. Fig. 4 confirms the trends found above. Errors for the ipsilateral region are about 3 dB smaller than errors for the contralateral region at $N = 1$, and differences between the two regions slowly decrease to approximately 1 dB at $N = 15$. Moreover, the errors for the ipsilateral region decrease almost monotonically, which is not the case for the contralateral region. While UP clearly performs worst, results for the alignment methods are comparable, except that FDTA produces larger errors for $N \leq 2$, and SUpDEq yields the lowest errors at $N = 1$, especially for the contralateral case. For $N \geq 3$, the differences between the methods diminish, and their performances become more and more similar.

To get a better impression of the spatial dependency of the spectral distortion, Fig. 5 shows $\Delta G(\Omega)$ for selected SH orders and SUpDEq. This shows that the region of large errors is generally small and quickly decreases with increasing SH order. For $N = 3$, frequency averaged errors above 3 dB are approximately found within a 45° radius cone around $\Omega = (270°, 0°)$, whereas the cone's radius

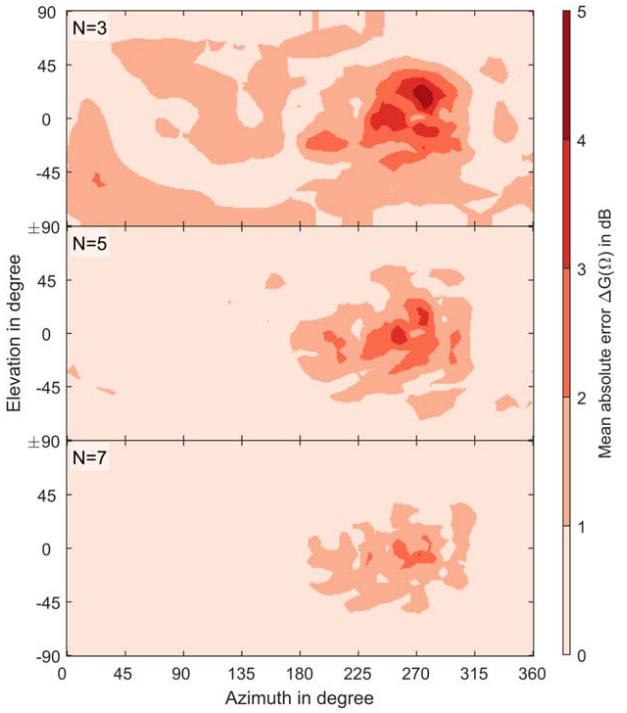


Fig. 5. Left ear energetic error $\Delta G(\Omega)$ for SUpDEq and selected SH orders N.

decreases to about 10° at $N = 7$. For comparison, the supplementary material provides similar plots for SH orders up to $N = 15$ and all methods [40, Figs. S6–S10], indicating similar behavior across the alignment methods.

4 PERCEPTUAL EVALUATION

The aim of the listening experiment was to determine the minimum required SH order N and thus the minimum required number of sampling points of a sparse HRTF set for which interpolated HRTFs are indistinguishable from the reference. To determine this so-called point of subjective equality (PSE), we implemented an adaptive ABX test, in which the SH order N of the sparse HRTF set is adapted according to the response of the subject. This was done for UP and SUpDEq as an example of the approaches discussed in Sec. 3 for three different test signals (noise, speech, and percussion) and two sound source positions (off-center frontal and lateral). We decided to test different source positions and audio content rather than different alignment methods because (first) the physical evaluation revealed that for $N \geq 3$ all methods perform very similarly in terms of TOA restoration and spectral distortion and (second) to limit the duration of the cognitively demanding ABX listening test. We hypothesized that SUpDEq processing generally leads to lower PSEs, that the test signal has a significant influence on the PSEs, and that the lateral sound source position leads to higher PSEs than the frontal position.

4.1 Participants

A total of 32 participants between 21 and 49 years of age ($M = 27.31$ years, $Mdn = 26$ years, $SD = 5.69$) took part

in the experiment for monetary remuneration of €15 per hour. Most of them were students in media technology or electrical engineering. Of those, 23 participants (72%) had already taken part in previous listening experiments and were thus familiar with the dynamic binaural reproduction system and the test environment. All participants had self-reported normal hearing.

4.2 Setup

The experiment was conducted in the sound insulated anechoic chamber of TH Köln, Köln, Germany. Participants were seated on an office chair with a mount holding a tablet computer at eye level about 0.50 m away in front of them, with which the responses were given. We used the MATLAB-based software Scale [45] to implement, control, and execute the experiment. For dynamic binaural rendering, we employed a customized version of the SoundScape Renderer [46], which is capable of loading Spatially Oriented Format for Acoustics (SOFA) files [47] with an arbitrary sampling grid. For three-degrees-of-freedom head tracking (yaw, pitch, and roll), we used a Polhemus Fastrack with 120 Hz update rate.

As digital-to-analog converter and headphone amplifier, we employed an RME Fireface UFX audio interface, and for playback, we used Sennheiser HD600 headphones. To minimize the influence of the headphones, we applied a generic headphone compensation filter, which was designed as a minimum phase finite impulse response filter with 2,048 taps using regularized inversion [48]. The playback level was adjusted to $L_{eq} = 65$ dB(A). The audio interface was set to a buffer size of 256 samples at a sampling rate of 48 kHz. With these settings, the measured overall latency of the system is about 37 ms [49], which is well below assessed thresholds of just detectable system latency of about 60–70 ms [50].

4.3 Stimuli

To obtain HRTFs for the listening test, the reference HRTFs (see Sec. 3.1) were subsampled to Gaussian grids of SH order $1 \leq N \leq 44$ using Eqs. (3) and (4). We chose the Gaussian grid because the order can be increased linearly. In a second step, UP and SUpDEq were used to interpolate HRTFs to a full-spherical spatial sampling grid with a resolution of 1° in horizontal direction and 5° in vertical direction. As with the physical evaluation in Sec. 3, we used an optimal radius of $r_0 = 9.19$ cm and the left and right ear position $\phi_e = [90^\circ, 270^\circ]$ and $\theta_e = [0^\circ, 0^\circ]$ for the spherical head model applied in SUpDEq.

We chose $\Omega = (330^\circ, 0^\circ)$ and $\Omega = (90^\circ, 0^\circ)$ as nominal sound source positions, first to examine PSEs for a frontal position that still contains at least small binaural cues and second to investigate the more critical lateral source position, which was shown to lead to significant artifacts at the contralateral ear, even with preprocessing (see also Sec. 3). As anechoic test signals, we employed a pink noise burst with a length of 0.75 s (including 10-ms cosine-squared onset/offset ramps), a male speech sample of a German sentence with a length of 1.5 s, and a castanet percussion sequence of 1.5 s length. The noise burst represents the most critical test signal with respect to coloration and localization, while speech and castanets are less critical due to the fluctuating spectral content and in the case of speech also due to the natural band limitation. However, percussion and speech signals are more relevant for real-life applications than noise[3].

4.4 Procedure

The experiment was based on an ABX test, that is, a three-interval/two-alternative forced choice (3I/2AFC) paradigm, combined with an adaptive one-up one-down staircase procedure (Chapter 3 and Chapter 5 of [51]). This simple and robust method [52] is free of restrictive assumptions, widely used in psychophysics, and was found to be a good choice to obtain the PSE [53] (i.e., the 50% point on the psychometric function also referred to as the threshold of recognition). Since perceptual differences between HRTFs interpolated from different SH orders are certainly not interval-scaled, more efficient maximum-likelihood procedures such as QUEST [54] could not be used.

According to the $3 \times 2 \times 2$ within-subjects factorial design with the factors *test signal* (noise, speech, and percussion), *method* (UP and SUpDEq), and *sound source position* ($\Omega = (330^\circ, 0^\circ)$ and $\Omega = (90^\circ, 0^\circ)$), each participant had to perform 12 runs. Following the ABX paradigm, a sequence of three intervals was presented at each trial, with X always being played second to ensure direct comparability between the stimuli (the actual playback order was therefore AXB). The middle interval (X) was randomly assigned to the reference HRTF set (A) or the sparse HRTF set (B), resulting in the four possible sequences AAB, BAA, ABB, or BBA. After the sequence was presented, participants had to report whether the first (A) or the third (B) interval was equal to the second (X) interval by pressing the corresponding button on the graphical user interface displayed on the tablet. The three buttons labeled A, X, and B were arranged on a horizontal line and flashed green when the corresponding interval was played. However, the X button was deactivated to prevent wrong entries. Participants could neither repeat a trial nor continue without giving an answer.

If the response was correct, the SH order of the sparse HRTFs was increased by one in the next trial and decreased by one otherwise. Each run started at $N = 1$ to provide clear perceptual differences to the participants. A run was terminated when 16 reversals occurred, where a reversal is defined as a point where a series of steps changes from increasing to decreasing the SH order or vice versa.

Before starting the experiment, participants were briefly introduced to dynamic binaural synthesis and were given instructions about the experimental procedure. They were encouraged to perform small head movements when they felt that this made them more sensitive to differences. To maintain differences between the two nominal source positions, they were additionally instructed to keep their main

[3] Static binaural renderings of the stimuli are part of the supplementary material [40].

Table 1. Mean PSEs across subjects and 95 % between-subjects confidence intervals (CIs) of the means for all tested conditions.

	$\Omega = (330°, 0°)$			$\Omega = (90°, 0°)$		
	Noise	Speech	Perc	Noise	Speech	Perc
Mean PSEs						
Unprocessed	18.27	12.49	13.28	24.29	19.37	18.97
SUpDEq	10.27	6.05	6.92	21.92	17.79	16.55
95% CIs						
Unprocessed	±1.73	±1.61	±1.18	±2.03	±1.42	±1.15
SUpDEq	±1.55	±1.12	±1.12	±1.98	±1.64	±1.90

line of vision straight ahead and were not allowed to rotate their body. The experimenter visually monitored the participants with a camera to ensure that they did not disregard the instructions. In order to get familiar with the setup and the test procedure, the participants had to do a short training session before the actual experiment, which consisted of two runs terminated when eight reversals occurred, one with the noise and one with the speech signal. In total, each session lasted for about 45 to 60 min, including the verbal instruction, the training session, and a break after half of the runs.

4.5 Data Analysis

To calculate the PSEs, the first reversal was omitted (Chapter 7 of [55]), and thus, the PSE estimate was calculated as the averaged N across the last 15 reversals. Visual inspection of the data and Shapiro–Wilk tests for normality, corrected for multiple hypothesis testing according to Hochberg [56], showed no considerable violations of normality (see also [40, Fig. S11]). We thus analyzed the determined PSEs using a three-way repeated measures ANOVA with Greenhouse–Geisser (GG) correction [57] and the within-subjects factors test signal, method, and sound source position. For a more detailed analysis, we conducted a nested GG-corrected repeated measures ANOVA as well as Hochberg-corrected paired t tests (two-tailed) at a 0.05 significance level.

4.6 Results

Table 1 lists the mean PSEs across subjects as well as the 95% between-subjects confidence intervals of the means for all tested conditions. The graphical overview of the data in Fig. 6 shows the interindividual variation in the determined PSEs (left panel) and the mean PSEs across subjects (right panel). The plots clearly support our three initial hypotheses, which are statistically confirmed by the ANOVA summarized in Table 2.

PSEs for SUpDEq are significantly lower than for UP resulting in a drastic decrease of the minimum number of measurement directions required to obtain SH interpolated HRTFs that are indistinguishable from the reference. The strong main effect of method revealed by the ANOVA statistically confirms this finding (Table 2, row M).

The sound source position has a strong influence on the PSEs. With both methods, the minimum required SH order increases significantly for the lateral source (90°, 0°) compared to the frontal (330°, 0°). The ANOVA yielded a strong main effect of source position (effect size $\eta_p^2 = 0.90$, see Table 2, row P) and thus statistically confirms the high perceptual relevance of the sound source position. Furthermore, the benefit of SUpDEq is smaller for the lateral position than for the frontal position, which is confirmed by the significant interaction effect between method and sound source position (Table 2, row M × P). Nevertheless, a nested ANOVA for the six lateral conditions showed a significant main effect of method suggesting that SUpDEq still provides improvements for the lateral position ($F(1,31) = 11.44$, $p = .002$, $\eta_p^2 = .27$, $\varepsilon = 1$).

Table 2. Results of the three-way repeated measures ANOVA with the within-subjects factor test signal (S), method (M), and sound source position (P).

Source	*df*	*F*	*MSE*	ε	η_p^2	*p*
S	2 , 62	48.59	19.90	1	.61	<.001*
M	1 , 31	101.48	19.37	1	.77	<.001*
P	1 , 31	272.92	26.02	1	.90	<.001*
S × M	2 , 62	.89	13.02	1	.03	.416
S × P	2 , 62	1.99	11.64	.97	.06	.147
M × P	1 , 31	58.45	9.49	1	.65	<.001*
S × M × P	2 , 62	.54	10.68	.92	.02	.573

ε, Greenhouse–Geisser (GG) epsilon; *p*, GG–corrected *p*-values. Note that GG correction is appropriate only for within-subject tests with more than one degree of freedom in the numerator.

Regardless of the method, the test signal has a strong influence on the PSEs, which is clearly demonstrated by the significant main effect of test signal revealed by the ANOVA (Table 2, row S). The speech and castanet signals require lower SH orders than the more critical noise signal. Paired t tests at each factor level of method and sound source position (e.g., Noise/UP/(330°, 0°) vs. Speech/UP/(330°, 0°)) confirmed that the PSEs for speech and castanets are always significantly lower than for noise (all $p < .001$). However, similar comparisons between speech and castanets showed no significant differences (all $p > .27$), indicating that both test signals are similarly critical.

5 DISCUSSION

5.1 Comparison Between Algorithms

The physical evaluation in Sec. 3 showed that HRTFs interpolated with the four investigated time-alignment methods are comparable in most cases. However, considerable differences were found in two cases. First, there are differences in the alignment and TOA restoration at low SH orders of $N \leq 2$. SUpDEq performs best in this case, presumably because it correctly models low-frequency phase effects involved in the diffraction around the sphere/head. Second, spectral differences at contralateral source positions remain up to SH orders of $N > 15$. In this region, the HRTF spectra exhibit fast changes across space, which requires higher SH orders for a physically correct interpolation. Caused by the insufficient SH order, aliasing errors

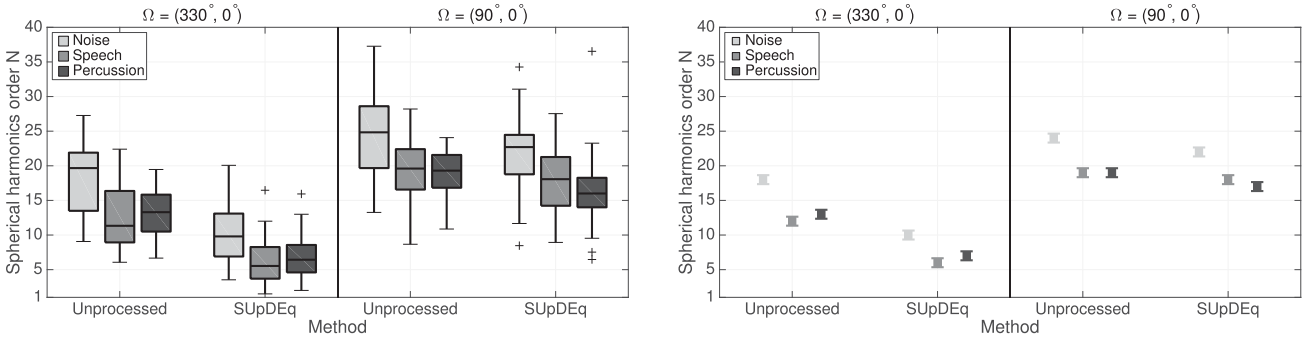


Fig. 6. Interindividual variation in the determined PSEs (left panel) and the mean PSEs across subjects (right panel) as a function of the method (abscissa), the test signal (shades of gray), and the sound source position (left or right half of each panel). The box plots (left panel) show the median and the (across participants) interquartile range (IQR) per condition; whiskers display 1.5 × IQR below the 25th or above the 75th percentile and outliers are indicated by plus signs. The error bars in the mean plots (right panel) display 95 % within-subjects confidence intervals [58, 59], based on the error term of the respective main effect of method.

occur that drastically differ between algorithms due to differences in the aligned magnitude *and* phase spectra.

5.2 Required SH Order for SUpDEq

The results of the listening experiment in Sec. 4 clearly show the advantages of time-alignment—SUpDEq, in this particular case—compared to SH interpolation of unprocessed HRTFs. Using SUpDEq, a sparse HRTF set with an SH order of $N \approx 7$ was sufficient for speech and castanet content presented from the front $\Omega = (330°, 0°)$ to achieve a binaural rendering that is indistinguishable from the reference. Without preprocessing, this requires an order of $N \approx 13$ and thus about three times more HRTFs ($14^2/8^2$). The pink noise signal presented from the front resulted in mean PSEs of $N \approx 10$ using SUpDEq and $N \approx 18$ without processing. In informal discussions after the experiment, the participants named high-frequency spectral differences as being the dominant cue for distinguishing the stimuli in the direct comparison. A broadband noise signal thus causes stronger perceptual differences and higher PSEs than band-limited speech and spectrally fluctuating castanets.

The lateral direction $\Omega = (90°, 0°)$ showed to be much more critical than the frontal direction. For speech and castanets, the mean PSEs were in the range $16 \leq N \leq 18$ for SUpDEq and $19 \leq N \leq 20$ for UP. For noise, the mean PSEs further increased to $N \approx 22$ using SUpDEq and $N \approx 24$ for the unprocessed case. The statistical analysis still showed a significant improvement with SUpDEq processing, but the benefits were much smaller than for frontal sound incidence.

Based on the physical evaluation in Sec. 3 and our previous study [16], we expected higher PSEs for lateral sound incidence due to the increased spectral distortions in the contralateral region. The distortion is caused by distinct magnitude interference patterns in the contralateral HRTF that change strongly even for small changes in the source position. This results in high SH orders in the contralateral region that cannot be reduced by means of time-alignment and causes sparsity errors in the interpolated HRTFs. Since the aliasing component of the sparsity error heavily depends on the sampling grid, these errors do not decrease monotonically with order, as can be seen in Figs. 3 and 4. As a result, the interpolated HRTFs show different interference patterns that are clearly distinguishable from the reference in a direct comparison, even more so with small head movements.

It should be kept in mind that the PSE is the most demanding quality criterion and that many applications do not require HRTFs that are indistinguishable from the reference. In the highly critical listening experiment, participants were able to suppress the nearly error-free signals at the ipsilateral "louder" ear and exploit spectral distortions at the contralateral "quieter" ear to distinguish between reference and SH-interpolated HRTFs. However, it is reasonable to assume that the perceived coloration is dominated by the "louder" ipsilateral ear and that spectral distortions at the contralateral ear are often less critical in reference-free listening.

In addition, the largest errors are contained in a narrow cone with a radius of approximately 10° already for an SH order of $N = 7$ (see Fig. 5). Because SUpDEq correctly models the ITD—the main localization cue in the horizontal plane [29]—already at order $N = 1$ (see Fig. 2), the left/right localization should not be problematic, even for lateral source positions. Although up/down localization relies on spectral cues [3], results from listening tests [23] and auditory modeling [19] suggest that an SH order of $N = 4$ maintains enough spectral detail for this task. Accordingly, coloration and localization, which are perhaps the two most important quality aspects besides the PSE, should be sufficiently good even for SH orders that are lower than the values determined with the present listening experiment.

Due to the similarity between the algorithms observed in the physical evaluation, it appears reasonable to assume that results obtained in the perceptual evaluation for SUpDEq also apply (approximately) to the other methods. However, more perceptual studies are required to generalize the results, and different thresholds might be found, especially for lateral sources.

5.3 Comparison to Previous Work

A comparison of our results with other studies is not directly possible because, to the best of our knowledge, there is no other study that has estimated PSEs for SH interpolated HRTFs. Using a 2AFC test, Pike and Tew [18] showed that SH-based HRTF interpolation with and without OBTA is indistinguishable from the reference at $N = 35$. In general, our results support the findings of Pike and Tew, even though one participant in our experiment achieved a PSE of $N \approx 37$ for the condition Noise/UP/(90°, 0°). However, the 95th percentile of this condition is $N \approx 32$, so it can be assumed that using $N = 35$ is sufficient for most listeners.

Using a MUSHRA test, Pike and Tew further showed that time-alignment of a sparse HRTF set with $N = 5$ reduces perceptual differences for a frontal source position, whereas a lateral source at $\Omega = (260°, 0°)$ still produces significant perceptual differences. This agrees with our analysis in Fig. 5, where the lateral source tested by Pike and Tew lies in the region of the largest spectral errors. It is also interesting to note that the frontal sources in the MUSHRA study of Pike and Tew received median quality ratings of about 90% in the case of time-aligned HRTF interpolation and a pink noise test signal. The fact that in the present experiment the median PSE for a similar condition was $N \approx 10$ further supports our assumption that quality-based listening experiments lead to lower minimum required SH orders of sparse HRTF sets.

5.4 Future Work

The physical and perceptual evaluation showed that spectral errors in the contralateral region remain the main challenge for time-alignment–based SH interpolation of HRTFs. Even if the phase components were perfectly eliminated, high SH orders were still necessary to describe the complex interference structure of the HRTF magnitude. To decrease the error in this region, (de-)equalization functions that approximate the HRTF better than the spherical head model used with SUpDEq might help to decrease the error in this region. Furthermore, a qualitative listening test to compare different alignment approaches would be interesting to assess the extent to which the differences discovered in the physical evaluation affect auditory perception.

6 CONCLUSION

In this paper, we performed a physical evaluation of four approaches for SH interpolation of time-aligned HRTFs and a perceptual evaluation of one selected time-alignment approach, namely the SUpDEq method. The systematic comparison showed the similarity of the different pre- and postprocessing techniques. For this reason, it is not surprising that the physical evaluation revealed that all methods perform similarly well in mitigating sparsity and reconstruction errors that occur in SH interpolation of unprocessed HRTFs. However, the analysis also showed that all discussed methods have drawbacks in the region around the contralateral ear.

The listening experiment showed the perceptual benefits of time-alignment on the example of the tested SUpDEq method. In all tested conditions, the minimum SH order required to achieve indistinguishability from a reference was significantly smaller than for SH interpolation without preprocessing. The results suggest that with an SH order of $N \approx 7$ (at least 64 measurement directions), interpolated HRTFs will be indistinguishable or close to indistinguishable from the reference for source positions in the vicinity of the median plane, while perceptual differences will be negligible for most remaining source positions and applications in spatial audio[4]. At order $N = 7$, the physical evaluation showed similar results for all tested methods. Thus, computationally less-demanding methods as PC and FDTA might be preferred in this case. However, differences in low-order processing still exist, and SUpDEq showed the lowest errors when using first-order HRTF sets.

7 ACKNOWLEDGMENT

This work was funded by the German Federal Ministry of Education and Research (BMBF) under the support code 03FH014IX5-NarDasS. We would like to thank Dr. Heinrich R. Liesefeld (LMU Munich, Department of Psychology) for his advice on the experimental design and the statistical analysis. We would also like to thank all participants of the listening experiment.

8 REFERENCES

[1] M. Vorländer, *Auralization* (Springer-Verlag, Berlin, Germany, 2008), http://doi.org/10.1007/978-3-540-48830-9.

[2] A. Roginska and P. Geluso, *Immersive Sound—The Art and Science of Binaural and Multi-Channel Audio* (Routledge, New York, NY, 2018), https://doi.org/10.4324/9781315707525.

[3] J. Blauert, *Spatial Hearing—The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1996).

[4] W. G. Gardner and D. M. Keith, "HRTF Measurements of a KEMAR," *J. Acoust. Soc. Am.*, vol. 97, no. 6, pp. 3907–3908 (1995 Jun.), https://doi.org/10.1121/1.412407.

[5] B. Bernschütz, "A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100," in *Proceedings of the 39th DAGA*, pp. 592–595 (2013).

[6] F. Brinkmann, A. Lindau, S. Weinzierl, S. Van De Par, M. Müller-Trapet, R. Opdam, and M. Vorländer, "A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations," *J. Audio Eng. Soc.*, vol. 65, no. 10, pp. 841–848 (2017 Oct.), https://doi.org/10.17743/jaes.2017.0033.

[7] V. R. Algazi, R. O. Duda, and D. M. Thompson, "The CIPIC HRTF Database," in *Proceedings of the IEEE Workshop on the Applications of Signal Pro-*

[4]Compare static binaural renderings provided in the supplementary material [40].

cessing to Audio and Acoustics, pp. 99–102 (2001 Oct.), https://doi.org/10.1109/ASPAA.2001.969552.

[8] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A Cross-Evaluated Database of Measured and Simulated HRTFs Including 3D Head Meshes, Anthropometric Features, and Headphone Impulse Responses," *J. Audio Eng. Soc.*, vol. 67, no. 9, pp. 705–718 (2019 Sep.), https://doi.org/10.17743/jaes.2019.0024.

[9] J.-G. Richter, *Fast Measurement of Individual Head-Related Transfer Functions*, doctoral dissertation, RWTH Aachen (2019), https://doi.org/10.30819/4906.

[10] W. Zhang, T. D. Abhayapala, R. A. Kennedy, and R. Duraiswami, "Insights Into Head-Related Transfer Function: Spatial Dimensionality and Continuous Representation," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2347–2357 (2010), https://doi.org/10.1121/1.3336399.

[11] Z. Ben-Hur, D. L. Alon, B. Rafaely, and R. Mehra, "Loudness Stability of Binaural Sound with Spherical Harmonic Representation of Sparse Head-Related Transfer Functions," *EURASIP J. Audio Speech Music Process.*, vol. 2019, no. 5, pp. 1–14 (2019), https://doi.org/10.1186/s13636-019-0148-x.

[12] B. Rafaely, "Analysis and Design of Spherical Microphone Arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143 (2005 Jan.), https://doi.org/10.1109/TSA.2004.839244.

[13] B. Bernschütz, *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording*, doctoral dissertation, TU Berlin (2016), http://dx.doi.org/10.14279/depositonce-5082.

[14] B. Bernschütz, A. V. Giner, C. Pörschmann, and J. M. Arend, "Binaural Reproduction of Plane Waves With Reduced Modal Order," *Acta Acust. united Ac.*, vol. 100, no. 5, pp. 972–983 (2014 Sep./Oct.), https://doi.org/10.3813/AAA.918777.

[15] M. Zaunschirm, C. Schoerkhuber, and R. Hoeldrich, "Binaural Rendering of Ambisonic Signals by HRIR Time Alignment and a Diffuseness Constraint," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3616–3627 (2018), https://doi.org/10.1121/1.5040489.

[16] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 6, pp. 1060–1071 (2019 Jun.), https://doi.org/10.1109/TASLP.2019.2908057.

[17] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Analyzing Head-Related Transfer Function Measurements using Surface Spherical Harmonics," *J. Acoust. Soc. Am.*, vol. 104, no. 4, pp. 2400–2411 (1998), https://doi.org/10.1121/1.423749.

[18] C. W. Pike, *Evaluating the Perceived Quality of Binaural Technology*, doctoral dissertation, University of York (2019).

[19] F. Brinkmann and S. Weinzierl, "Comparison of Head-Related Transfer Functions Pre-Processing Techniques for Spherical Harmonics Decomposition," presented at the *AES International Conference on Audio for Virtual and Augmented Reality* (2018 Aug.), conference paper P9-3.

[20] 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, no. 12, pp. 2249–2262 (2019 Dec.), https://doi.org/10.1109/TASLP.2019.2945479.

[21] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in Proceedings of the *44th DAGA*, pp. 339–342 (2018).

[22] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording* (Springer, Cham, Switzerland, 2019), https://doi.org/10.1007/978-3-030-17207-7.

[23] G. D. Romigh, D. S. Brungart, R. M. Stern, and B. D. Simpson, "Efficient Real Spherical Harmonic Representation of Head-Related Transfer Functions," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 921–930 (2015 Aug.), https://doi.org/10.1109/JSTSP.2015.2421876.

[24] G. Dagan, N. R. Shabtai, and B. Rafaely, "Spatial Release from Masking for Binaural Reproduction of Speech in Noise with Varying Spherical Harmonics Order," *Appl. Acoust.*, vol. 156, pp. 258–261 (2019 Dec.), https://doi.org/10.1016/j.apacoust.2019.07.015.

[25] B. Rafaely, *Fundamentals of Spherical Array Processing* (Springer-Verlag, Berlin, Germany, 2015), https://doi.org/10.1007/978-3-662-45664-4.

[26] J. Ahrens, *Analytic Methods of Sound Field Synthesis* (Springer-Verlag, Berlin, Germany, 2012), https://doi.org/10.1007/978-3-642-25743-8.

[27] A. Andreopoulou and B. F. G. Katz, "Identification of Perceptually Relevant Methods of Inter-Aural Time Difference Estimation," *J. Acoust. Soc. Am.*, vol. 142, no. 2, pp. 588–598 (2017), https://doi.org/10.1121/1.4996457.

[28] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the Unity Delay," *IEEE Sig. Proc. Mag.*, vol. 13, no. 1, pp. 30–60 (1996 Jan.), https://doi.org/10.1109/79.482137.

[29] F. L. Wightman and D. J. Kistler, "The Dominant Role of Low-Frequency Interaural Time Differences in Sound Localization," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1648–1661 (1992), https://doi.org/10.1121/1.402445.

[30] C. Pörschmann and J. M. Arend, "Obtaining Dense HRTF Sets from Sparse Measurements in Reverberant Environments," presented at the *AES International Conference on Immersive and Interactive Audio* (2019 Mar.), conference paper 15.

[31] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Spatial Upsampling of Individual Sparse Head-Related Transfer Function Sets by Directional Equalization," in *Proceedings of the 23rd International Congress on Acoustics*, pp. 4870–4877 (2019), http://doi.org/10.18154/RWTH-CONV-239484.

[32] J. M. Arend and C. Pörschmann, "Synthesis of Near-Field HRTFs by Directional Equalization of Far-Field

Datasets," in *Proceedings of the 45th DAGA*, pp. 1454–1457 (2019).

[33] J. M. Arend and C. Pörschmann, "Spatial Upsampling of Sparse Head-Related Transfer Function sets by Directional Equalization—Influence of the Spherical Sampling Scheme," in *Proceedings of the 23rd International Congress on Acoustics*, pp. 2643–2650 (2019), http://doi.org/10.18154/RWTH-CONV-238939.

[34] V. Tourbabin and B. Rafaely, "On the Consistent Use of Space and Time Conventions in Array Processing," *Acta Acust. united Ac.*, vol. 101, no. 3, pp. 470–473 (2015 May/Jun.), https://doi.org/10.3813/AAA.918843.

[35] B. Rafaely, B. Weiss, and E. Bachmat, "Spatial Aliasing in Spherical Microphone Arrays," *IEEE Trans. Audio Speech Lang. Process.*, vol. 55, no. 3, pp. 1003–1010 (2007 Mar.), https://doi.org/10.1109/TSP.2006.888896.

[36] R. Baumgartner, P. Majdak, and B. Laback, "Modeling Sound-Source Localization in Sagittal Planes for Human Listeners," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 791–802 (2014), https://doi.org/10.1121/1.4887447.

[37] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a Spherical-Head Model from Anthropometry," *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 472–479 (2001 Jun.).

[38] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. (Pearson Higher Education, Inc., Upper Saddle River, NJ, 2010).

[39] V. Benichoux, M. Rébillat, and R. Brette, "On the Variation of Interaural Time Differences with Frequency," *J. Acoust. Soc. Am.*, vol. 139, no. 4, pp. 1810–1821 (2016), https://doi.org/10.1121/1.4944638.

[40] J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions – Supplementary Material," (2021), https://doi.org/10.5281/zenodo.4289971.

[41] J. E. Mossop and J. F. Culling, "Lateralization for Large Interaural Delays," *J. Acoust. Soc. Am.*, vol. 104, no. 3, pp. 1574–1579 (1998), https://doi.org/10.1121/1.424369.

[42] C. Hold, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev, "Improving Binaural Ambisonics Decoding by Spherical Harmonics Domain Tapering and Coloration Compensation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 261–265 (2019), https://doi.org/10.1109/ICASSP.2019.8683751.

[43] M. Slaney, "Auditory Toolbox: A Matlab Toolbox for Auditory Modeling Work - Technical Report #1998-010" (1998).

[44] B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl, "SOFiA Sound Field Analysis Toolbox," in *Proceedings of the International Conference on Spatial Audio (ICSA)*, pp. 8–16 (2011).

[45] A. Vazquez Giner, "Scale—Conducting Psychoacoustic Experiments with Dynamic Binaural Synthesis," in *Proceedings of the 41st DAGA*, pp. 1128–1130 (2015).

[46] M. Geier, J. Ahrens, and S. Spors, "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods," presented at the *124th Convention of the Audio Engineering Society* (2008 May), convention paper 7330.

[47] Audio Engineering Society, "AES69-2015: AES Standard for File Exchange—Spatial Acoustic Data File Format" (2015).

[48] V. Erbes, M. Geier, H. Wierstorf, and S. Spors, "Free Database of Low-Frequency Corrected Head-Related Transfer Functions and Headphone Compensation Filters," presented at the *127th Convention of the Audio Engineering Society* (2017 May), eBrief 325.

[49] J. M. Arend, T. Lübeck, and C. Pörschmann, "A Reactive Virtual Acoustic Environment for Interactive Immersive Audio," presented at the *2019 AES International Conference on Immersive and Interactive Audio* (2019 Mar.), conference paper 9.

[50] A. Lindau, "The Perception of System Latency in Dynamic Binaural Synthesis," in *Proceedings of the 35th DAGA*, pp. 1063–1066 (2009).

[51] F. A. A. Kingdom and N. Prins, *Psychophysics: A Practical Introduction*, 1st ed. (Academic Press, London, United Kingdom, 2009), https://doi.org/10.1016/C2012-0-01278-1.

[52] H. Levitt, "Transformed Up-Down Methods in Psychoacoustics," *J. Acoust. Soc. Am.*, vol. 49, no. 2B, pp. 467–477 (1971), https://doi.org/10.1121/1.1912375.

[53] T. S. Meese, "Using the Standard Staircase to Measure the Point of Subjective Equality: A Guide Based on Computer Simulations," *Perc. Psychophys.*, vol. 57, no. 3, pp. 267–281 (1995), https://doi.org/10.3758/bf03213053.

[54] A. B. Watson and D. G. Pelli, "QUEST: A Bayesian Adaptive Psychometric Method," *Perc. Psychophys.*, vol. 33, no. 2, pp. 113–120 (1983), https://doi.org/10.3758/bf03202828.

[55] S. A. Gelfand, *Hearing—An Introduction to Psychological and Physiological Acoustics*, 6th ed. (CRC Press, Boca Raton, FL, 2017).

[56] Y. Hochberg, "A Sharper Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, vol. 75, no. 4, pp. 800–802 (1988 Dec.), https://doi.org/10.1093/biomet/75.4.800.

[57] G. V. Glass, P. D. Peckham, and J. R. Sanders, "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance," *Rev. Educ. Res.*, vol. 42, no. 3, pp. 237–288 (1972), https://doi.org/10.3102/00346543042003237.

[58] G. R. Loftus and M. E. J. Masson, "Using Confidence Intervals in Within-Subject Designs," *Psychon. Bull. Rev.*, vol. 1, no. 4, pp. 476–490 (1994), https://doi.org/10.3758/bf03210951.

[59] J. Jarmasz and J. G. Hollands, "Confidence Intervals in Repeated-Measures Designs: The Number of Observations Principle." *Can. J. Exp. Psychol.*, vol. 63, no. 2, pp. 124–138 (2009), https://doi.org/10.1037/a0014164.

THE AUTHORS


Johannes M. Arend


Fabian Brinkmann


Christoph Pörschmann

Johannes M. Arend received a B.Eng. degree in media technology from HS Düsseldorf (Germany) in 2011 and an M.Sc. degree in media technology from TH Köln, Köln, Germany, in 2014. Since 2015, he has been a Research Fellow and working toward a Ph.D. at TH Köln, Köln, Germany, and TU Berlin, Berlin, Germany, in the field of spatial audio with a focus on binaural technology, auditory perception, and audio signal processing.

•

Fabian Brinkmann received an M.A. degree in communication sciences and technical acoustics in 2011 and a Dr. rer. nat. degree in 2019 from the Technical University of Berlin, Berlin, Germany. He focuses on the fields of signal processing and evaluation approaches for spatial audio.

•

Christoph Pörschmann studied Electrical Engineering at the Ruhr-Universität Bochum, Bochum, Germany, and Uppsala Universitet, Uppsala, Sweden. In 2001, he obtained his Dr.-Ing. degree from the Electrical Engineering and Information Technology Faculty of the Ruhr-Universität Bochum as a result of his research at the Institute of Communication Acoustics. Since 2004, he has been Professor of Acoustics at TH Köln, Köln, Germany. His research interests are in the field of virtual acoustics, spatial hearing, and the related perceptual processes.

2.2 Comparison of Spherical Harmonics and Nearest-Neighbor based Interpolation of Head-Related Transfer Functions

Pörschmann, C., Arend, J. M., Bau, D., & Lübeck, T. (2020). In *Proc. of the AES International Conference on Audio for Virtual and Augmented Reality (AVAR), Redmond, WA, USA* (pp. 1–10).

Audio Engineering Society

Conference Paper

Presented at the International Conference on
Audio for Virtual and Augmented Reality
2020 August 17 – 19 , Online

Comparison of Spherical Harmonics and Nearest-Neighbor based Interpolation of Head-Related Transfer Functions

Christoph Pörschmann[1], Johannes M. Arend[1,2], David Bau[1,2], and Tim Lübeck[1,2]

[1] *Institute of Communications Engineering, TH Köln - University of Applied Sciences, Cologne 50679, Germany*
[2] *Audio Communication Group, Technical University of Berlin, Berlin 10587, Germany*

Correspondence should be addressed to Christoph Pörschmann (Christoph.Poerschmann@th-koeln.de)

ABSTRACT

Spatial upsampling of head-related transfer functions (HRTFs) measured on a sparse grid is an important issue, particularly relevant when capturing individual datasets. While early studies mostly used nearest-neighbor approaches, ongoing research focuses on interpolation in the spherical harmonics (SH) domain. The interpolation can either be performed on the complex spectrum or separately on magnitude and unwrapped phase. Furthermore, preprocessing methods can be applied to reduce the spatial complexity of the HRTF dataset before interpolation. We compare different methods for the interpolation of HRTFs and show that SH and nearest-neighbor based approaches perform comparably. While generally a separate interpolation of magnitude and unwrapped phase outperforms an interpolation of the complex spectra, this can be compensated by appropriate preprocessing methods.

1 Introduction

For a spatial presentation of sound sources in virtual acoustic environments (VAEs), monaural and binaural cues, which are mainly caused by the shape of the pinna and the head, need to be considered. While spectral information is the main cue to determine elevation, we use differences between the signals reaching the left and the right ear for lateral localization. These binaural differences manifest in interaural time differences (ITDs) and interaural level differences (ILDs). In many headphone-based VAEs, head-related transfer functions (HRTFs) are applied to describe the sound incidence from a source, which is typically in the far-field, to the left and right ear incorporating both, monaural and binaural cues [1, 2]. For many VAEs, HRTF sets are used, which describe the sound reaching a listener from various directions and which can be captured, e.g. on a spherical sampling grid.

To adequately represent the spatial cues for all incident directions, a large number of HRTFs needs to be determined. These HRTF sets can either be measured with a dummy head (e.g. [3, 4, 5]), or individual HRTFs can be obtained with specialized measurement setups for high fidelity spatial audio (e.g. [6, 7, 8]). According to Lindau and Weinzierl [9] for a noise stimulus, which can be regarded as the most critical one, a resolution of 2° in the horizontal and 1° in the vertical plane is required for an artifact-free dynamic auralization if no interpolation of the HRTFs is applied. This results in an extremely large number of 32,000 HRTFs on an equiangular angle grid, which for most applications needs to be reduced by applying physically and

perceptually suitable interpolation methods. Already in 1993, Wenzel and Foster [10] investigated nearest-neighbor based approaches for the interpolation of non-individual HRTFs. The study showed that localization accuracy was largely unaffected by interpolation even for large intervals both in azimuth and elevation. In the studies of Hartung et al. [11] and Djelani et al. [12], first attempts were taken to improve the interpolation by an appropriate preprocessing. The authors suggested to determine the initial time delay of the head-related impulse responses (HRIR, the time-domain equivalent of an HRTF) from the position of the absolute maximum and to remove the initial time delay from the HRIRs. Then the interpolation of the spatially closest impulse responses was performed linearly, and finally, an interpolated initial time delay was added. Another approach to improve the interpolation is to interpolate not directly the complex spectrum of the HRTFs, but magnitude and phase separately. In this context, the results of Hartung et al. [11] indicated that a separate interpolation of the magnitude and unwrapped phase of the HRTF performed better than an interpolation of the complex spectrum, both technically and perceptually. This was confirmed in a study of Langendijk and Bronkhorst [13], who analyzed the perceptual influence of the spatial resolution of measured HRTFs. All these studies applied the neighboring directions for determining the interpolated directions. For the weighting of the contributing directions, different methods were used, e.g. spline-based interpolation [11] distance-weighting [12] or natural-neighbor interpolation [14].

Alternatively, the interpolation can be carried out in the spherical harmonics (SH) domain [15, ch. 6], [16, ch. 1]. Here, the HRTF set, measured on a spherical grid, is decomposed into spherical basis functions of different orders N, whereby the number of measured directions directly corresponds to the maximum stable resolvable spatial order N. In this case, HRTF interpolation is performed by evaluating the SH coefficients at the corresponding interpolated directions. Already in 1998, Evans et al. [17] proposed an SH-based interpolation method which was either carried out in time or frequency domain. Moreover, the authors investigated how removing of the ITDs, which can be regarded as a preprocessing step, could further enhance their approach. As the SH basis functions form a spatially continuous set of solutions of the wave equation, an interpolation in the SH domain yields a physically correct and spatially continuous HRTF representation as long as $N \geq kr$, with $k = \frac{\omega}{c}$, ω the frequency and c the propagation velocity of sound [18, 19]. Therefore, a minimum spatial order of $N = 32$ and accordingly, depending on the sampling scheme, at least 1089 measurements are required for artifact-free interpolation over the entire audio bandwidth up to 20 kHz, when assuming $r = 8.75$ cm as the average human head radius [20] and the speed of sound $c = 343$ m/s. Consequently, the interpolation of sparse HRTF sets, i.e., HRTF sets which are measured with a low spatial resolution (e.g. for 38 sampling points on a Lebedev grid [18]) results in an incomplete description of the spatial and spectral properties and leads to order-limitation artifacts affecting high-frequency components and binaural cues [19, 21, 22]. In the same way, as described above, the SH-based interpolation can separately be applied to the magnitude and phase spectra. As shown e.g. by Romigh et al. [23] and Brinkmann and Weinzierl [24], this separation can reduce the spatial order and accordingly lead to reduced errors caused by the interpolation.

Furthermore, time-aligning the HRTFs in preprocessing before SH transform [25, 21, 24] can reduce order-limitation artifacts. In this context, we recently presented the SUpDEq (Spatial Upsampling by Directional Equalization) method [26] and proposed a spatial equalization of HRTFs with the corresponding rigid sphere transfer functions (STFs). The STFs can be regarded as a dataset that features basic temporal and magnitude spectral components but omits information about the head's fine structure. The spatial equalization removes both the linear phase component of the HRTFs as well as other typical amplitude-related features caused e.g. by diffraction. The results show a significant reduction of the spectral errors and a refined representation of the binaural cues in the interpolated HRTF set. A similar approach has been recently published in [27].

As many influencing factors affect the interpolation performance, a direct comparison of the different interpolation approaches is very challenging. For example, there are no studies directly comparing the performance of SH-based interpolation to nearest-neighbor approaches. Furthermore, the impact of the preprocessing depends on other influencing factors, e.g., the interpolation method or the representation of the HRTF set on which the interpolation is performed. This paper aims to analyze and evaluate some of these interactions and find an optimal combination of the different influencing factors. We compare SH-based interpolation to

a nearest-neighbor approach, and either directly interpolate the complex spectrum, or alternatively magnitude and unwrapped phase separately. As an example of for a nearest-neighbor approach, we chose the natural-neighbor interpolation [14]. Subsequently, we compare the results of the interpolation with and without the SUpDEq method, as an example for preprocessing that removes the initial delays. This paper is organized as follows. In Section 2, we describe the synthesis of the datasets applying the varying influencing factors for HRTF interpolation. In Section 3, we compare the different approaches regarding their deviations from high-resolution reference dataset, followed by a discussion in Section 4. Finally, in Section 5, we conclude and describe in which way the examined interpolation and preprocessing methods can be optimally combined.

2 Methods

2.1 Influencing Factors

Interpolation Method The first influencing factor relates to the interpolation method itself. When applying SH interpolation, the sampling scheme of the dataset defines the maximal resolvable order N of the SH representation. In a previous study [28], we examined how different spherical sampling schemes affect spatial upsampling of HRTFs when using either common SH interpolation or the SUpDEq method. The results revealed that different sampling schemes, e.g. equiangular, Gaussian, Lebedev, or Fliege at the same spatial order N only marginally affect the SH upsampling performance. If specific areas are not covered by measurements, e.g. as for individual HRTF sets in which often areas in the lower hemisphere are missing, regularization approaches as according to Tikhonov [29] can be applied, which are not further discussed or investigated here though. However, it should be mentioned that it is a specific property of SH interpolation that by transforming the sampled functions to the SH domain, errors or inaccuracies of one measured point on the sphere unavoidable affect the entire SH representation.

This is completely different for nearest-neighbor based approaches. Here, deviations to a reference, caused, for example, by inaccuracies in the measurement setup or procedure, only influence the area around these sampling positions. Other areas of the sphere remain unaffected. The performance of the nearest-neighbor approaches might further depend on the method used for weighting the neighboring sampling points for the interpolation. However, in pilot studies investigating the influence of the weighting method, we could only find a slight influence on the performance, which is why we applied the natural-neighbor interpolation [14] as one possible nearest-neighbor approach. For the implementation we applied the Sound Field Synthesis Toolbox [30], which calculates the weights for the interpolation based on Voronoi diagrams.

Data representation As a second factor, we vary the representation of the data on which we perform the interpolation. We either apply the interpolation to the complex spectra of the HRTFs, often simply denoted as linear interpolation, or to the magnitude and phase separately. Fig. 1 shows the result for a simple example of either SH-based or nearest-neighbor-based interpolation between two HRIRs with different amplitudes and different initial delays. It can be observed that the interpolation of complex spectra results in two peaks. This probably disturbs the ITD and causes, among others, spectral deviations. Alternatively, magnitude and phase can be separated before interpolation. As shown in Fig. 1 this results in a single peak of the HRIR, which is averaged in time and amplitude. This can be regarded as an appropriate result of the interpolation, both in terms of the temporal and magnitude spectral structure.

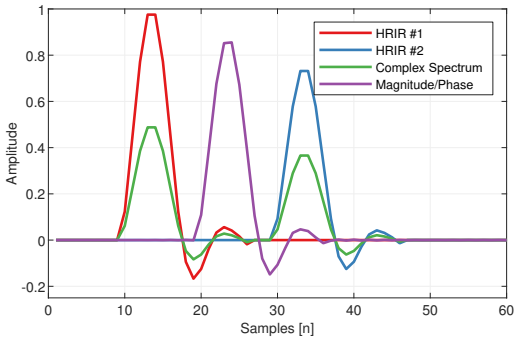


Fig. 1: Left ear signals of two simplified HRIRs (red, blue), interpolation of the complex spectrum (green), and interpolation of magnitude and unwrapped phase separately (purple).

Preprocessing Finally, as a third influencing factor, we examine the benefit of preprocessing the HRTF sets. The basic idea of many of those preprocessing methods is to remove the temporal [11, 12, 25, 21, 24, 27] or the temporal and magnitude spectral differences

[26] from the HRTF set prior to the interpolation. Especially the compensation of the initial delay of the HRIRs is extremely important in this context, as it can reduce the artifacts of interpolating the complex spectra, as shown in Fig. 1. As an example of such a preprocessing method, we perform spatial upsampling using the SUpDEq method, which has been described in detail in [26].The basic idea is as follows: The sparse HRTF set is equalized by spectral division with the corresponding directional rigid sphere transfer functions (hereinafter called equalization dataset), yielding an equalized sparse HRTF dataset. Generally speaking, the equalization dataset represents a simplified HRTF set only featuring the basic shape of a spherical head, but without any information on the specific shape of the outer ear. As the equalization reduces the differences between neighboring sampling points on the sparse grid and accordingly limits the spatial complexity of the dataset, it decreases errors during the subsequent spatial upsampling. After spatial upsampling, a de-equalized dataset is obtained by a directional multiplication with a set of rigid sphere transfer functions according to the dense grid.

2.2 Materials

Although interpolation methods are typically used to reduce the effort of measuring individual HRTF sets, in this study we applied HRTFs of a Neumann KU100 dummy head as the reference set $HRTF_{ref}$ [4], providing an accurately measured HRTF set with high spatial density as ground truth data. The set was measured on a Lebedev grid with 2702 sampling points and can be used for SH processing up to $N = 44$. From this reference, we generated several sparse datasets: Lebedev grids with 14, 38, 86, and 170 sampling points corresponding to a maximum spatial order of $N = 2, 4, 7, 10$, respectively. The sparse sets were obtained from the SH representation of the reference dataset by applying the inverse SH transform on the respective grids. Then, for each of the sparse grids, we created different types of datasets by upsampling to a Lebedev grid with 2702 sampling points using different spatial upsampling methods. We applied an SH-based interpolation of the complex spectrum, in the following denoted as $HRTF_{SH}$, SH-based interpolation performed separately on magnitude and unwrapped phase denoted as $HRTF_{SH,PH}$, natural-neighbor interpolation of the complex spectrum denoted as $HRTF_{NAT}$, and separated for magnitude and unwrapped phase denoted as $HRTF_{NAT,PH}$. For each of these sets we created two variants, the first without any preprocessing, the second applying the SUpDEq method, hereinafter referred to as *DEQ*.

Table 1: Overview of the different HRTF sets created for evaluation of the interpolation.

Dataset	Interpolation method	Data represent.
$HRTF_{SH}$	spherical harmonics	complex spectrum
$HRTF_{SH,PH}$	spherical harmonics	magnitude / phase
$HRTF_{NAT}$	natural neighbor	complex spectrum
$HRTF_{NAT,PH}$	natural neighbor	magnitude / phase

3 Technical Evaluation

In a first step, we compare the reference to the different test datasets for an exemplary direction. For this, we determined left ear HRTFs for an azimuth of $\phi = 45°$ and an elevation of $\theta = 0°$, based on a sparse HRTF set sampled on a Lebedev grid with 38 sampling points allowing for an SH processing up to $N = 4$. Fig. 2 shows the magnitude and impulse responses resulting from the different interpolation methods. The interpolation of the complex spectra leads to strong deviations above 3 kHz due to spatial aliasing (see e.g. [19, 22, 21]). For this incidence direction, the natural-neighbor interpolation (NAT) performs worst and shows deviations of 5 dB from the reference already at 3 kHz. In contrast, apart from some slight ripples, the preprocessed datasets are in good agreement with the reference. The same holds for all datasets which are based on a separate interpolation of magnitude and phase. Thus, it seems that even without preprocessing, adequate results can be obtained. Generally, Fig. 2 represents typical results for a source in the frontal area. For this exemplary direction the closest sampling points of the sparse grid are located at an angular distance of 11°.

Next we analyze the spectral deviations from the reference set as a function of N on a Lebedev grid with 2702 sampling points as test sampling grid. For this, the frequency-dependent spectral differences per sampling point were calculated in dB as

$$\Delta G(\omega, \Omega_t) = 20lg \frac{|\, HRTF_{ref}(\omega, \Omega_t)\,|}{|\, HRTF_{test}(\omega, \Omega_t)\,|}, \qquad (1)$$

where $HRTF_{ref}$ is the HRTF extracted from the reference set measured on a dense grid and $HRTF_{test}$ is the

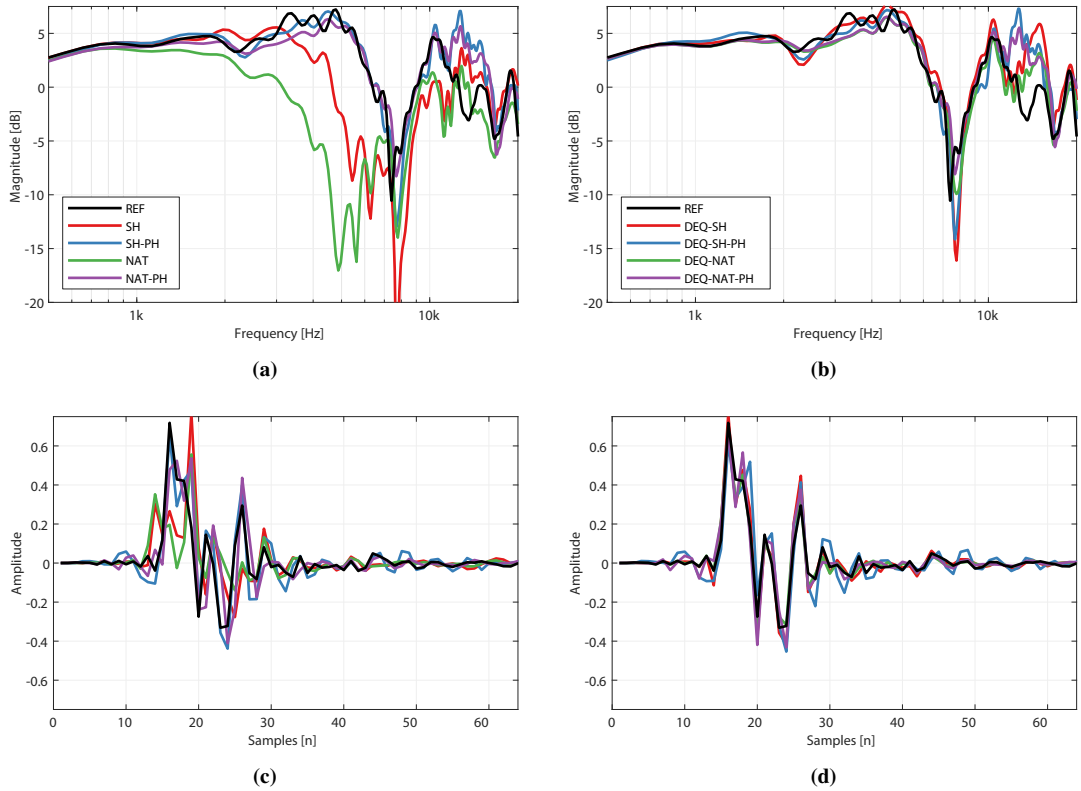


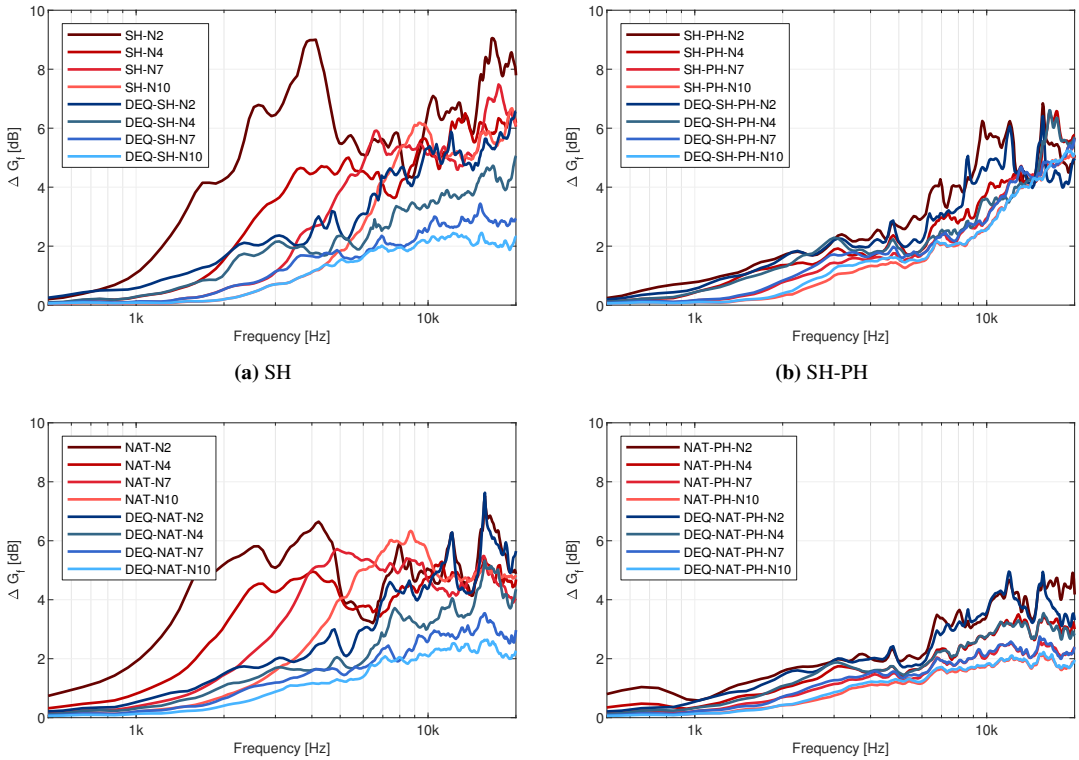
Fig. 2: Left ear magnitudes (a,b) and impulse responses (c,d), extracted from the reference set $HRTF_{ref}$ (black) and the different interpolated sets for an incidence direction of $\phi = 45°$ (azimuth), $\theta = 0°$ (elevation). SH interpolation of the complex spectrum $HRTF_{SH}$ (red), SH interpolation separately on magnitude and unwrapped phase $HRTF_{SH,PH}$ (blue), natural-neighbor interpolation of the complex spectrum $HRTF_{NAT}$ (green), natural-neighbor interpolation separately on magnitude and unwrapped phase $HRTF_{NAT,PH}$ (purple). The plots show interpolation results of a sparse HRTF set with 38 sampling points. Left column (a,c): No preprocessing, right column (b,d): SUpDEq preprocessing.

HRTF calculated from the upsampled dataset, both at the direction Ω_t. Then, the absolute value of $\Delta G(\omega, \Omega_t)$ was averaged over all sampling points Ω_t to obtain the frequency-dependent measure $\Delta G_f(\omega)$ (in dB):

$$\Delta G_f(\omega) = \frac{1}{n_{\Omega_t}} \sum_{\Omega_t=1}^{n_{\Omega_t}} |\, \Delta G(\omega, \Omega_t)\,|, \qquad (2)$$

Fig. 3 presents the spectral differences $\Delta G_f(\omega)$ for Lebedev grids with 14, 38, 86, and 170 sampling points. The plot shows that for interpolation performed on the complex spectrum ($HRTF_{SH}$, $HRTF_{NAT}$), the spectral differences are significantly smaller when applying SUpDEq preprocessing (left column). In contrast, when performing the interpolation separately for magnitude and unwrapped phase ($HRTF_{SH,PH}$, $HRTF_{NAT,PH}$), $\Delta G_f(\omega)$ is even without preprocessing significantly smaller compared to an interpolation of the complex spectra (right column). In this case, the SUpDEq method has nearly no impact on the spectral deviations. The smallest deviations can be observed for $HRTF_{NAT,PH}$, which are already for the smallest grid with only 14 sampling points below 4 dB up to 10 kHz.

Concluding the analysis of the spectral differences, we analyze the spatial distribution of the deviations. For this, we calculate the absolute value of $\Delta G(\omega, \Omega_t)$ averaged across frequency ω to obtain one value $\Delta G_{sp}(\Omega_t)$ (in dB) per sampling point:



(a) SH **(b)** SH-PH

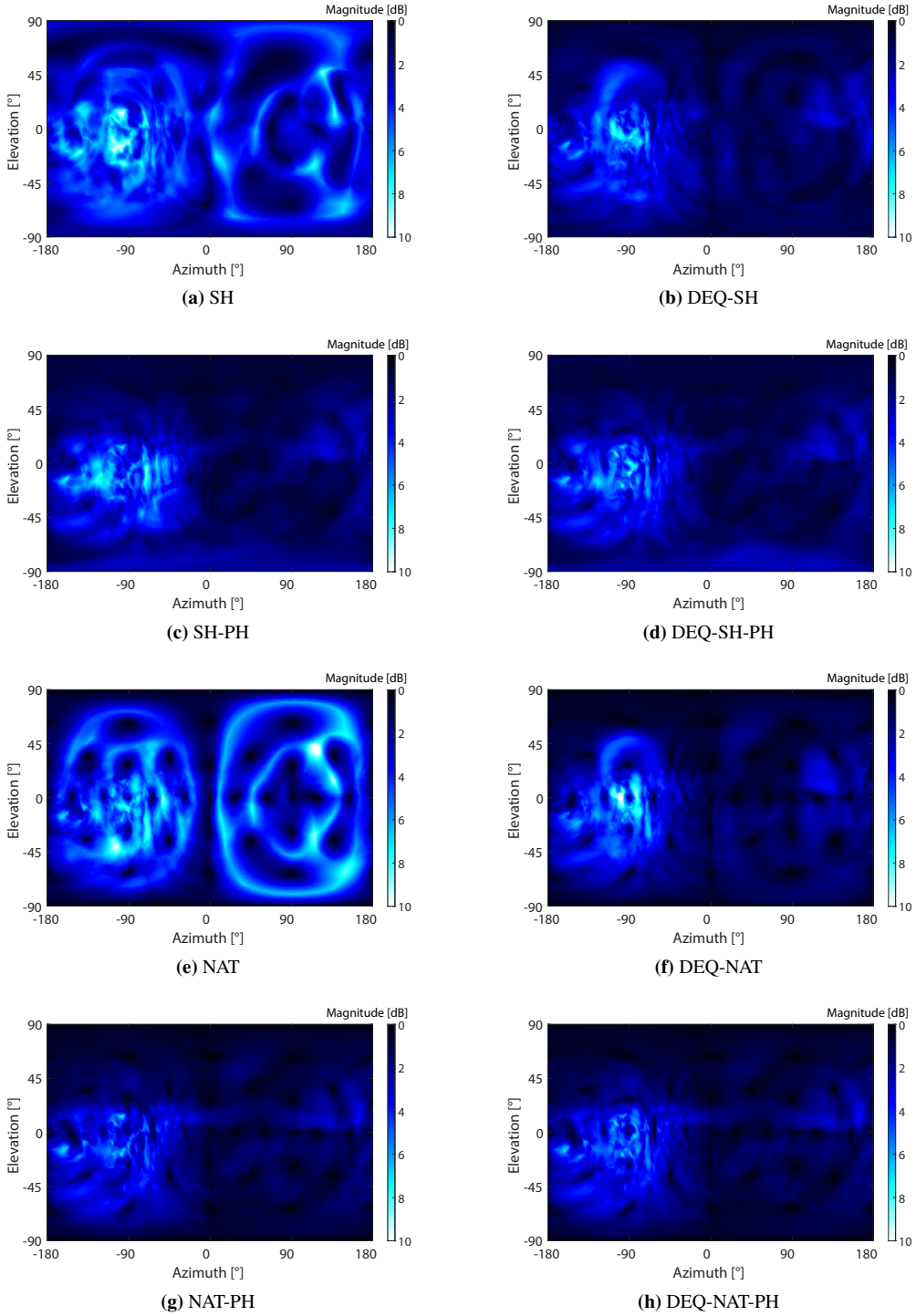
(c) NAT **(d)** NAT-PH

Fig. 3: Spectral differences $\Delta G_f(\omega)$ between the reference set and interpolated datasets (left ear). No preprocessing (red), SUpDEq preprocessing (blue). (a) $HRTF_{SH}$, (b) $HRTF_{SH.PH}$, (c) $HRTF_{NAT}$, (d) $HRTF_{NAT,PH}$

$$\Delta G_{sp}(\Omega_t) = \frac{1}{n_\omega} \sum_{\omega=1}^{n_\omega} | \Delta G(\omega, \Omega_t) | . \qquad (3)$$

Fig. 4 shows the spectral differences $\Delta G_{sp}(\Omega_t)$ for the different interpolation approaches for a sparse Lebedev grid with 38 sampling positions for $f \leq 10\,\text{kHz}$. The test sampling grid $\Omega_t = \{(\phi_1, \theta_1), \ldots, (\phi_T, \theta_T)\}$ was generated for ϕ and θ in steps of 1° over the entire surface of the sphere. From the plots, it can be observed that the spectral differences are maximal for contralateral directions, mostly because diffraction around the head influences the sound incidence, resulting in an increased spatial complexity at the contralateral ear. Please refer to [26] for a more detailed discussion on these effects. Furthermore, independent of the interpolation method, the spectral deviations $\Delta G_{sp}(\Omega_t)$ are very similar when applying SUpDEq (Fig. 4 b,d,f,h). Without preprocessing, spectral differences are significantly higher when interpolating the complex spectra (Fig. 4 a,e) and lower when performed separately for magnitude and unwrapped phase (Fig. 4 c,g).

Finally, we compare the ILDs and ITDs of the reference HRTF set and the interpolated sets, which are based on a dataset with 38 sampling points on a Lebedev grid. For this, we extracted HRTFs in the horizontal plane in steps of 1° from the reference and the interpolated sets and calculated broadband ILDs as the ratio between the energy of the left and right ear HRIR. To determine the ITDs, we applied a threshold-based onset detection (-10 dB relative to the peak magnitude value) on the ten times up-sampled and low-pass filtered HRIRs (10th-order Butterworth low-pass at 3 kHz). This value was chosen because the auditory system mainly considers the ITD for frequencies below 1.5 kHz [1]. Fig. 5 shows the respective ILDs and ITDs based on a Lebedev grid with 38 sampling points. For the datasets without preprocessing (a,c), the SH-based interpolation of the complex magnitude results



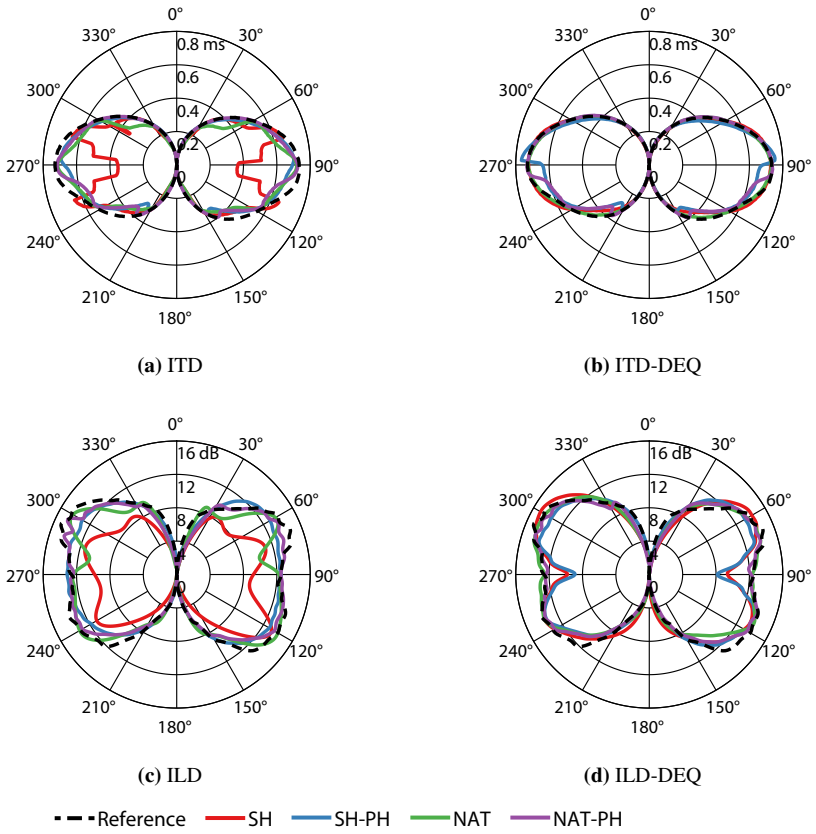
(a) SH **(b)** DEQ-SH

(c) SH-PH **(d)** DEQ-SH-PH

(e) NAT **(f)** DEQ-NAT

(g) NAT-PH **(h)** DEQ-NAT-PH

Fig. 4: Spectral differences $\Delta G_{sp}(\Omega_t)$ per sampling point for $HRTF_{SH}$ (a,b), for $HRTF_{SH,PH}$ (c,d), $HRTF_{NAT}$ (e,f) and for $HRTF_{NAT,PH}$ (g,h) for a Lebedev grid with 38 sampling points and $f \leq 10\,\text{kHz}$. The left column (a,d,e,g) shows the results without preprocessing, the right column (b,d,f,h) with SUpDEq preprocessing.



(a) ITD **(b)** ITD-DEQ

(c) ILD **(d)** ILD-DEQ

Fig. 5: ILDs (a,b) and ITDs (c,d) in the horizontal plane for the reference (black) HRTF set and for $HRTF_{SH}$ (red), $HRTF_{SH,PH}$ (blue), $HRTF_{NAT}$ (green) $HRTF_{NAT,PH}$ (purple) interpolated sets based on a Lebedev grid with 38 sampling points. In the left column (a,c) the sets without preprocessing are shown, in the right column (b,d) the SUpDEq-processed sets. The angle represents the azimuth ϕ of the sound source. The radius describes the magnitude of the level differences (in dB) or time differences (in ms).

in relevant errors both for ITDs and ILDs. For example, in the frontal region ($-60° \leq \phi \leq 60°$), the ITD error reaches values of up to $100\,\mu s$ and the ILD error reaches $6\,dB$. When interpolating magnitude and unwrapped phase separately (NAT-PH and SH-PH), binaural cues are reconstructed much better with ITD errors below $20\,\mu s$ and ILD errors below $2\,dB$ for the frontal region ($-60° \leq \phi \leq 60°$). For the datasets processed with the SUpDEq method (b,d), all ITDs and ILDs match the reference quite well. For example, for NAT, the ITD and the ILD errors are less than $20\,\mu s$ and $2\,dB$ respectively for all incidence directions in the horizontal plane. The ITD errors are below the JND, which according to Klockgether and van de Par [31] are about $20\,\mu s$ in anechoic environments. The ILD deviations are above the JNDs, for which in this study values between $0.5\,dB$ and $1.5\,dB$ depending on stimulus and room properties were determined.

4 Discussion

The temporal and magnitude spectral structure of an exemplary HRTF, as well as spectral deviations and binaural cues are deteriorated the most if the complex spectra is interpolated and no preprocessing is applied. Preprocessing, which in this case reduces the spectral complexity of the HRTF set, results in significantly smaller errors, and the spectral differences vary only slightly between the methods and mainly depend on the resolution of the sampling grid. Even though only

tested informally, it can be expected that other preprocessing approaches, e.g. [25, 21, 24, 27], lead to similar results. Generally speaking, an ideal preprocessing would completely remove the phase information. Interpolating the complex spectra of such preprocessed HRTFs is comparable to an interpolation of the magnitude spectra separated from the phase. Thus, good results can be obtained when interpolating magnitude and unwrapped phase separately. For example, for a Lebedev grid with 86 sampling points (allowing for stable SH processing up to $N = 7$), the spectral differences $\Delta G_f(\omega)$ remain below 3 dB for frequencies up to 10 kHz. In this case, it is of minor importance which interpolation method was used or whether preprocessing was applied.

Generally, we found the lowest deviations to the reference when performing nearest-neighbor based interpolation of magnitude and unwrapped phase. This holds for the temporal structure, the spectral deviations, and the binaural cues. The use of a preprocessing method is nearly obsolete in this case, although ITDs and ILDs for lateral sound incidence might marginally benefit from preprocessing.

5 Conclusion

In this study, we compared different methods for HRTF interpolation. The results show that a separate interpolation of magnitude and unwrapped phase performs much better than interpolating the complex spectra. While for sampling grids with a very low spatial resolution (e.g. 38 sampling positions on a Lebedev grid), the nearest-neighbor based approach performs best, for grids with higher spatial resolution SH-based interpolation and nearest-neighbor based approaches become more and more comparable. Furthermore, we showed that appropriate preprocessing, removing the initial delays of the HRTFs can be applied in combination with all tested interpolation methods. In this case, interpolating the complex spectra shows comparable results as a separate interpolation of magnitude and unwrapped phase. The preprocessing has some small impact when interpolating magnitude and phase separately as it slightly improves the binaural cues for lateral sound incidence.

References

[1] Blauert, J., *Spatial Hearing - The Psychophysics of Human Sound Localization*, MIT Press, Cambridge, MA, revised edition, 1996.

[2] Vorländer, M., *Auralization*, Springer-Verlag, Berlin Heidelberg, 2008.

[3] Gardner, W. G. and Martin, K. D., "HRTF measurements of a KEMAR," *J. Acoust. Soc. Amer.*, 97(6), pp. 3907–3908, 1995.

[4] Bernschütz, B., "A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100," in *Proceedings of the 39th DAGA*, pp. 592–595, 2013.

[5] Arend, J. M., Neidhardt, A., and Pörschmann, C., "Measurement and Perceptual Evaluation of a Spherical Near-Field HRTF Set," *Proceedings of the 29th Tonmeistertagung - VDT International Convention*, pp. 52–55, 2016.

[6] Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B., "Head-Related Transfer Functions of Human Subjects," *J. Audio Eng. Soc*, 43(5), pp. 300–321, 1995.

[7] Algazi, V., Duda, R. O., Thompson, D. M., and Avendano, C., "The CIPIC HRTF database," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, October, pp. 99–102, 2001.

[8] Bomhardt, R. and Fels, J., "Mismatch between Interaural Level Differences Derived from Human Heads and Spherical Models," in *Proceedings of the 140th AES Convention*, pp. 1–10, 2016.

[9] Lindau, A. and Weinzierl, S., "On the Spatial Resolution of Virtual Acoustic Environments for Head Movements in Horizontal, Vertical and Lateral Direction," in *Proceedings of the EAA Symposium on Auralization*, pp. 1–6, 2009.

[10] Wenzel, E. and Foster, S., "Perceptual consequences of interpolating head-related transfer functions during spatial synthesis," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 102–105, 1993.

[11] Hartung, K., Braasch, J., and Sterbing, S. J., "Comparison of different methods for the interpolation of head-related transfer functions," in *AES 16th International Conference: Spatial Sound Reproduction*, pp. 319–329, 1999.

[12] Djelani, T., Pörschmann, C., Sahrhage, J., and Blauert, J., "An Interactive Virtual-Environment Generator for Psychoacoustic Research II: Collection of Head-Related Impulse Responses and Evaluation of Auditory Localization," *Acta Acustica United Acustica*, 86(6), pp. 1046–1053, 2000.

[13] Langendijk, E. H. A. and Bronkhorst, A. W., "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *J. Acoust. Soc. Amer.*, 107(1), pp. 528–537, 2000.

[14] Sibson, R., "A brief description of natural neighbor interpolation," in V. Barnett, editor, *Interpolating Multivariate Data*, chapter 2, pp. 21–36, John Wiley, 1981.

[15] Williams, E. G., *Fourier Acoustics - Sound Radiation and Nearfield Acoustical Holography*, Academic Press, London, UK, 1999.

[16] Rafaely, B., *Fundamentals of Spherical Array Processing*, Springer-Verlag, Berlin Heidelberg, 2015.

[17] Evans, M. J., Angus, J. A. S., and Tew, A. I., "Analyzing head-related transfer function measurements using surface spherical harmonics," *J. Acoust. Soc. Amer.*, 104(4), pp. 2400–2411, 1998.

[18] Rafaely, B., "Analysis and Design of Spherical Microphone Arrays," *IEEE Trans. Speech Audio Process.*, 13(1), pp. 135–143, 2005.

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

[20] Hartley, R. V. L. and Fry, T. C., "The Binaural Location of Pure Tones," *Physical Review*, 18(6), pp. 431–442, 1921.

[21] Zaunschirm, R., Markus, Schörkhuber, C., and Höldrich, "Binaural rendering of Ambisonic signals by HRIR time alignment and a diffuseness constraint," *J. Acoust. Soc. Amer.*, 143(6), pp. 3616 – 3627, 2018.

[22] Ben-Hur, Z., Brinkmann, F., Sheaffer, J., Weinzierl, S., and Rafaely, B., "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *J. Acoust. Soc. Amer.*, 141(6), pp. 4087–4096, 2017.

[23] Romigh, G. D., Brungart, D. S., Stern, R. M., and Simpson, B. D., "Efficient Real Spherical Harmonic Representation of Head-Related Transfer Functions," *IEEE Journal on Selected Topics in Signal Processing*, 9(5), 2015.

[24] Brinkmann, F. and Weinzierl, S., "Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition," in *Proceedings of the AES Int. Conf. on Audio for Virtual and Augmented Reality*, pp. 1–10, 2018.

[25] Pike, C. and Tew, A., "Subjective Assessment of HRTF Interpolation with Spherical Harmonics," in *Proceedings of ICSA 2017*, 2017.

[26] Pörschmann, C., Arend, J. M., and Brinkmann, F., "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 27(6), pp. 1060 – 1071, 2019.

[27] Ben-Hur, Z., Alon, D. L., Mehra, R., and Rafaely, B., "Efficient Representation and Sparse Sampling of Head-Related Transfer Functions Using Phase-Correction Based on Ear Alignment," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 27(12), pp. 2249–2262, 2019.

[28] Arend, J. M. and Pörschmann, C., "Spatial upsampling of sparse head-related transfer function sets by directional equalization - Influence of the spherical sampling scheme," in *Proceedings of the 23rd International Congress on Acoustics*, 2019.

[29] Zotkin, D. N., Duraiswami, R., and Gumerov, N. A., "Regularized HRTF fitting using spherical harmonics," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, November, pp. 257–260, 2009.

[30] Wierstorf, H., Winter, F., Rettberg, T., Erbes, V., Hahn, N., Schultz, F., Hold, C., and Spors, S., "SFS Toolbox 2.5.0," 2019, doi:10.5281/zenodo.2597212.

[31] Klockgether, S. and van de Par, S., "Just noticeable differences of spatial cues in echoic and anechoic acoustical environments," *J. Acoust. Soc. Amer.*, 140(4), pp. EL352–EL357, 2016.

2.3 Spatial Upsampling of Sparse Head-Related Transfer Function Sets by Directional Equalization – Influence of the Spherical Sampling Scheme

Arend, J. M., & Pörschmann, C. (2019). In *Proc. of the 23rd International Congress on Acoustics (ICA), Aachen, Germany* (pp. 2643–2650). https://doi.org/10.18154/RWTH-CONV-238939

PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Spatial upsampling of sparse head-related transfer function sets by directional equalization - Influence of the spherical sampling scheme

Johannes M. AREND[(1)(2)†], Christoph PÖRSCHMANN[(1)]

[(1)]Institute of Communications Engineering, TH Köln, D-50679 Cologne, Germany,
[(2)]Audio Communication Group, TU Berlin, D-10587 Berlin, Germany
[†]Corresponding author, E-mail: johannes.arend@th-koeln.de

Abstract

Many immersive audio applications rely on a dense set of head-related transfer functions (HRTFs). However, often only measurements on a specific sparse grid are available. To obtain dense HRTF sets from sparse measurements, one common approach is to apply spatial interpolation in the spherical harmonics (SH) domain. However, the SH representation of sparse HRTF sets is order-limited, leading to spatial aliasing and truncation errors. In a recent publication, we presented the so-called SUpDEq method (Spatial Upsampling by Directional Equalization) for spatial upsampling of sparse HRTF sets. The approach is based on a directional equalization of the sparse set prior to the spherical Fourier transform to remove direction-dependent temporal and spectral components. This significantly reduces the spatial complexity of the sparse set, allowing for an enhanced interpolation at reduced SH orders. In this study we investigate how different spherical sampling schemes affect the performance of common SH interpolation and the SUpDEq method. For this, we compare spatially upsampled HRTF sets originally based on sparse equiangular, Gaussian, Lebedev, and Fliege grids at various SH orders to a reference. The influence of the different grids are assessed spectrally, temporally, and with localization models.

Keywords: Head-Related Transfer Functions (HRTFs), Spatial Audio, Spherical Harmonics

1 INTRODUCTION

Human sound source localization is based on binaural cues, i.e. interaural time differences (ITDs) and interaural level differenes (ILDs) between both ears, as well as on monaural cues, i.e. spectral distortions of the incoming sound caused mainly by the listener's pinna, head, and torso. Head-related transfer functions (HRTFs) contain these binaural and monaural cues and thus describe the sound incidence from a source to both ears [5].

For headphone-based virtual acoustic environments (VAEs), a set of HRTFs is essential. Ideally, such a set should include individual HRTFs for a large number of directions, typically measured on a sphere around a listener. However, measuring so-called dense sets of individual HRTFs requires special equipment, experience in handling the equipment and, depending on the measurement approach, can also be time-consuming (see e.g. [6]). For this reason, it seems appealing to measure only a small number of HRTFs on a sparse spherical sampling grid with a simplified measurement system, and to apply a specific interpolation or spatial upsampling method afterwards to generate a dense HRTF set with perhaps thousands of directions.

One popular approach for spatial upsampling is interpolation in the spherical harmonics (SH) domain. For this, an HRTF set captured on a spherical sampling scheme (also simply called spatial grid) is first transformed to the SH domain applying the spherical Fourier transform (SFT). The resulting spatially continuous representation of the HRTF set in the SH domain allows for interpolation, i.e. an HRTF for any desired direction can be obtained by means of the respective inverse spherical Fourier transform (ISFT) [13]. However, the SH representation and interpolation of sparse HRTF sets suffers from so-called sparsity errors, which is a combination of spatial aliasing and truncation errors [3]. For this reason, various pre- and post-processing methods have been proposed to reduce the sparsity error and thus to improve SH interpolation of sparse HRTF sets (see e.g. [7]).

Within this scope, we presented the SUpDEq method (SUpDEq - Spatial Upsampling by Directional Equalization) as a pre- and post-processing approach allowing improved SH interpolation of sparse HRTF sets [11].

In the respective paper, we examined the performance of the SUpDEq method regarding spectral and temporal features as well as concerning modeled localization performance of reconstructed HRTFs and showed that the approach clearly outperforms common SH interpolation in terms of these features. However, as the analysis was based only on the Lebedev sampling scheme [9], the present paper now provides further evaluation investigating the influence of various (sparse) spherical sampling schemes on the performance of the SUpDEq method. As the aliasing error strongly depends on the sampling scheme [3], this evaluation is of particular interest to ensure the general applicability of the proposed upsampling method.

In this paper, we therefore compare spatially upsampled HRTF sets originally based on sparse equiangular, Gaussian, Lebedev, and Fliege grids [13, Ch. 3][9][8] at various spatial orders N to a dense reference HRTF set. Similar to the evaluation in [11], we assess the impact of the grids on the spatially upsampled HRTF sets spectrally, temporally, and by means of localization models. To anticipate some of the results, the analysis showed that the sampling scheme has very little, if any, influence on the performance of the SUpDEq method.

2 SPHERICAL SAMPLING SCHEMES

A set of HRTFs is commonly measured at discrete points on a surrounding sphere according to a spherical sampling scheme. Such a full-spherical HRTF set can be transformed to the SH domain with the discrete SFT. The sampling schemes investigated in this paper provide closed-form expressions to calculate SH coefficients, whereas SH coefficients for arbitrary sampling configurations can be computed by an inversion of the respective SH matrix [13, Ch. 3]. The latter is however not further discussed here. Given a spherical sampling scheme L with a closed-form expression, the spherical HRTF set $H(\omega,\Omega_q)$ for the left and right ear (indices for left and right are omitted here and in the following for ease of display) can be described in the SH domain by the SH coefficients $h_{nm}(\omega)$ that are computed with the discrete SFT [13, p. 58]

$$\widehat{h}_{nm}(\omega) = \sum_{q=1}^{Q_L} \beta_q \; H(\omega,\Omega_q)[Y_n^m(\Omega_q)]^*, \tag{1}$$

with the temporal frequency ω, the Q_L directions $\Omega_q = \{(\phi_1,\theta_1),\ldots,(\phi_Q,\theta_Q)\}$ at azimuth ϕ and elevation θ, and the sampling weights β_q depending on the sampling scheme L. The notation $(\cdot)^*$ denotes complex conjugation and Y_n^m are the complex SH functions of order n and degree m. The ISFT can be applied to recover $H(\omega)$ at arbitrary angles allowing for SH interpolation [13, p. 17]

$$\widehat{H}(\omega,\Omega) = \sum_{n=0}^{N} \sum_{m=-n}^{n} h_{nm}(\omega) Y_n^m(\Omega)\,, \tag{2}$$

where N is the spatial order (also referred to as SH order). As discrete sampling of a function with infinite order induces spatial aliasing and truncation errors, the SH coefficients are only error-free up to a specific scheme-dependent order N_L. If the function sampled on the sphere is strictly order-limited, a sampling scheme providing a sufficient order N_L results in $h_{nm}(\omega) = \widehat{h}_{nm}(\omega)$. Similar, $H(\omega) = \widehat{H}(\omega)$ holds if N is chosen appropriately.

The maximum resolvable order of the sampling scheme N_L is generally defined by the number of directions (or sampling points) Q_L and by the way the sampling points are distributed around the surface of the sphere. This relationship can be expressed by $Q_L \geq \eta(N+1)^2$, with η describing the efficiency of the sampling scheme [12]. The SH order of an HRTF set however increases as the frequency increases, following the relation $N \sim kr$, with k the wavenumber and r the radius of a sphere surrounding the head [11][3]. Assuming an average human head radius of $r = 8.75\,cm$, a minimum SH order $N = 32$ is required to perform a nearly perfect SFT, ISFT, and thus SH interpolation of HRTFs for frequencies up to $20\,kHz$.

In research, various schemes have been developed in order to sample the sphere with the highest possible accuracy and efficiency. A good overview on different sampling approaches in the context of spatial audio can be found for example in [13, Ch. 3]. For this study, we focused on four different frequently applied schemes, namely the equiangular, Gaussian, Lebedev, and Fliege grids. The equiangular grids have a uniform distribution

of samples along ϕ and θ, with both angles sampled at $2(N+1)$ locations, requiring $4(N+1)^2$ samples in total [12]. The Gaussian grids require only $2(N+1)^2$ samples, as the elevation θ is only sampled at $(N+1)$ locations, resulting in a nearly-uniform distribution of samples along both angles [12]. However, the equiangular and the Gaussian sampling schemes do not provide uniform distributions of sample points on the surface of the sphere. The Lebedev and Fliege schemes however offer nearly-uniform distribution of samples around the surface of the sphere, with the advantage that even less sample points are required to reach a specific SH order. Thus, the Lebedev grids require approximately $1.3(N+1)^2$ samples whereas the Fliege grids only require $(N+1)^2$ sample points [12][13, Ch. 3]. Figure 1 shows the four introduced grids on the sphere, exemplarily of SH order $N=7$, resulting in 256 points for the equiangular grid (a), 128 points for the Gaussian grid (b), 86 points for the Lebedev grid (c), and 64 points for the Fliege grid (d).

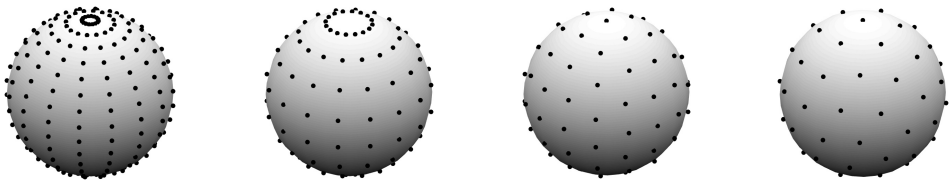


Figure 1. Equiangular (a), Gaussian (b), Lebedev (c), and Fliege (d) sampling schemes of SH order $N=7$.

3 SPATIAL UPSAMPLING BY DIRECTIONAL EQUALIZATION (SUpDEq)

The following section gives a brief overview of the basic concept behind the SUpDEq method, as illustrated in the block diagram in Fig. 2. Further details on the implementation and evaluation can be found in [11]. Basically, the approach aims at enhanced SH interpolation and spatial upsampling of sparse HRTF sets. To achieve this, a sparse HRTF set $H(\omega,\Omega_S)$ measured at S sampling points $\Omega_s=\{(\phi_1,\theta_1),\ldots,(\phi_S,\theta_S)\}$ is equalized directionally by spectral division with an appropriate equalization dataset $D_{EQ}(\omega,\Omega_s)$ before the SFT:

$$H_{EQ}(\omega,\Omega_s)=H(\omega,\Omega_s)/D_{EQ}(\omega,\Omega_s)\,. \tag{3}$$

As a rather good and established approximation of a human head, direction-dependent rigid sphere transfer functions for an incident plane wave [13, p. 44] are used as the equalization dataset. The spherical head model should match the respective human head as best as possible. As a first and easy to implement approach, the radius of the sphere is calculated according to the physical dimensions of the head [1] and the ears are positioned at $\phi=\pm 90^\circ$ and $\theta=0^\circ$ on the sphere. Using a spherical head model also has the advantage that it can be described analytically, which allows the calculation of the rigid sphere transfer functions at high SH orders $N_{high}\geq 32$. The directional equalization described in Eq. (3) significantly reduces the spatial complexity of the sparse HRTF set, therefore minimizing the required SH order for the SFT. The reason for the decrease of the SH order is that on the one hand, the equalization leads to a time-alignment of the HRTFs, similar to a re-centering, and on the other hand, direction-dependent influences of the sphere or the head are compensated.

After equalization, the equalized HRTF set $H_{EQ}(\omega,\Omega_s)$ is transformed to the SH domain with the SFT (Eq. (1)) at a low SH order N_{low} according to the maximum resolvable SH order of the spare sampling scheme. Then, an upsampled (equalized) HRTF set $\widehat{H}_{HRTF,EQ}$ is calculated on a dense sampling grid $\Omega_d=\{(\phi_1,\theta_1),\ldots,(\phi_D,\theta_D)\}$, with $D\gg S$ using the ISFT (Eq. (2)). Finally, HRTFs are reconstructed with a subsequent de-equalization by spectral multiplication with an appropriate de-equalization dataset D_{DEQ}:

$$\widehat{H}_{DEQ}(\omega,\Omega_d)=\widehat{H}_{EQ}(\omega,\Omega_d)\cdot D_{DEQ}(\omega,\Omega_d)\,. \tag{4}$$

Again, rigid sphere transfer functions for an incident plane wave can are used as the de-equalization dataset. In general, the de-equalization recovers energies at higher SH orders that were transformed to lower orders by

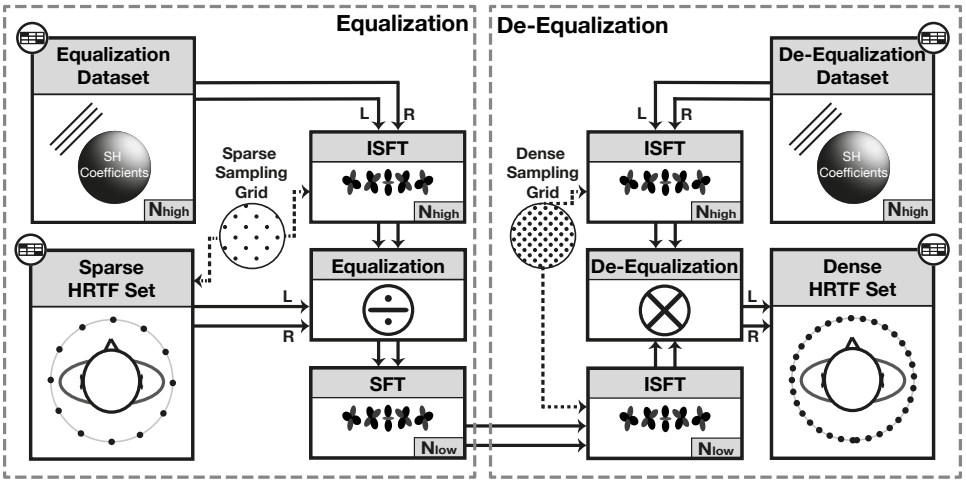


Figure 2. Block diagram of the SUpDEq method. Left panel: A sparse HRTF set is equalized on the corresponding sparse sampling grid and then transformed to the SH domain with $N = N_{low}$. Right panel: The equalized set is de-equalized on a dense sampling grid, resulting in a dense HRTF set.

the equalization. Similar as described in Sec. 2, $H = \widehat{H}_{DEQ}$ holds if, in this case, N_{low} is sufficient for the SFT of H_{EQ} and N_{high} is chosen appropriately. Otherwise, spatial aliasing and truncation errors occur, resulting in $H \approx \widehat{H}_{DEQ}$. The following section now analyzes the performance of the SUpDEq method with respect to the sparse sampling scheme underlying the input HRTF set in comparison to common SH interpolation.

4 INFLUENCE OF THE SAMPLING SCHEME

In our previous publication [11], we investigated the performance of the SUpDEq method for two different dummy heads, but only for Lebedev grids of different SH orders. To further ensure the general applicability of the SUpDEq method, the present paper now focuses on the influence of the spherical sampling scheme underlying the sparse input HRTF set. As a reference set, we used HRTFs of a Neumann KU100 dummy head that were measured on a Lebedev grid with 2702 sampling points [4]. This reference HRTF set was transformed to SH domain at $N = 35$, further referred to as $h_{REF,nm}$. The various sparse HRTF sets required as input data were generated by spatial subsampling of the reference set $h_{REF,nm}$ to the respective sparse equiangular, Gaussian, Lebedev, or Fliege grids of (limited) SH orders $N = 1-15$ applying the ISFT. Next, these sparse HRTF sets were spatially upsampled to a dense sampling grid (again the Lebedev grid with 2702 sampling points, further abbreviated Lebedev_{2702}), applying the SUpDEq method as well as (order-limited) SH interpolation without any pre- or post-processing before or after the SFT/ISFT. The upsampled dense HRTF sets were then again transformed to SH domain at $N = 35$, resulting in SH coefficients further referred to as $h_{DEQ,nm}$ and $h_{OL,nm}$, with DEQ standing for de-equalized and OL for (strictly) order-limited. The de-equalized and order-limited HRTFs, as hereinafter referred to, were then obtained via the ISFT at the direction required for the respective analysis method. The optimal radius for the rigid sphere model used for (de-)equalization was calculated based on the dimensions of the Neumann KU100 dummy head [1], leading to r = 9.19 cm.

4.1 Spectral differences

As a first error measure, we analyzed the spectral differences between $h_{REF,nm}$ and $h_{DEQ,nm}$ or $h_{OL,nm}$ as a function of the SH order N on various test sampling grids with T sampling points $\Omega_t = \{(\phi_1, \theta_1), \ldots, (\phi_T, \theta_T)\}$.

The frequency-dependent spectral differences per sampling point were calculated in dB as

$$\Delta g(\omega,\Omega_t) = 20lg\frac{|\,H_{REF}(\omega,\Omega_t)\,|}{|\,H_{TEST}(\omega,\Omega_t)\,|}, \tag{5}$$

where H_{REF} is the left ear HRTF extracted from $h_{REF,nm}$ and H_{TEST} is the left ear HRTF extracted from $h_{OL,nm}$ or $h_{DEQ,nm}$ at the sampling point Ω_t. Then, the absolute value of $\Delta g(\omega,\Omega_t)$ was averaged across all sampling points Ω_t to obtain the frequency-dependent measure $\Delta G_f(\omega)$ (in dB)

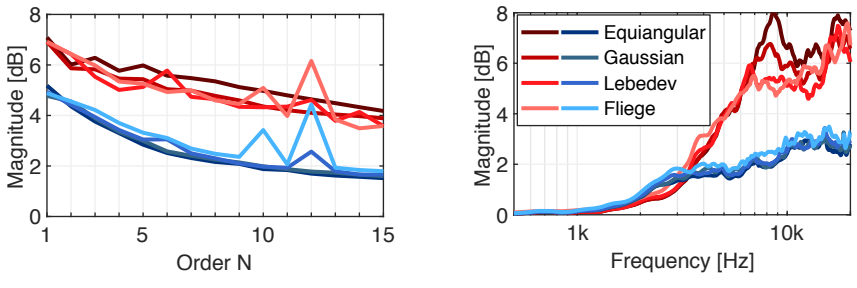
$$\Delta G_f(\omega) = \frac{1}{n_{\Omega_t}}\sum_{\Omega_t=1}^{n_{\Omega_t}} |\,\Delta g(\omega,\Omega_t)\,|, \tag{6}$$

and across ω and Ω_t, resulting in a single value ΔG (in dB) describing the spectral difference

$$\Delta G = \frac{1}{n_{\Omega_t}}\frac{1}{n_\omega}\sum_{\Omega_t=1}^{n_{\Omega_t}}\sum_{\omega=1}^{n_\omega} |\,\Delta g(\omega,\Omega_t)\,|\,. \tag{7}$$

Figure 3 (a) shows the spectral differences ΔG across N over the full audio bandwidth for the four different sampling schemes applying the SUpDEq method or order-limited interpolation. The test sampling grid Ω_t was the reference Lebedev$_{2702}$ grid. Independent of the sampling scheme, SUpDEq processing results in about 2 dB less spectral differences than order-limited interpolation. The Fliege scheme however has distinct outliers at $N = 10$ and $N = 12$ for both upsampling methods. Interestingly, exactly at these orders, some of the calculated weights are negative, which is something Fliege and Maier could not explain [8]. Applying the SFT according to Eq. (1), the negative weights most probably lead to a phase shift in the complex SH coefficients, certainly resulting in reconstruction errors when transformed back with the ISFT. Apart from these outliers, the spectral differences for the four different sampling schemes are pretty similar, indicating that the performance of the SUpDEq method is independent of the sampling scheme. For order-limited interpolation, the equiangular scheme leads to slightly higher spectral differences than the other schemes.

Figure 3 (b) illustrates the frequency-dependent spectral differences, exemplarily at $N = 7$, for the four different sampling schemes applying the SUpDEq method or order-limited interpolation. As before, Ω_t was the Lebedev$_{2702}$ grid. It can be seen that the spectral differences are significantly smaller for the SUpDEq method than for order-limited interpolation, and furthermore that order-limited interpolation leads to a sharp increase in spectral differences above the spatial aliasing frequency. Regarding the SUpDEq method, the Fliege scheme performs a little worse than the three other schemes, but overall there is only a marginal influence of the sampling scheme on the performance of the method. Furthermore, the equiangular scheme induces slightly higher spectral differences than the other schemes when applying order-limited interpolation.



(a) (b)

Figure 3. Spectral differences in dB (left ear) between reference HRTF set and order-limited (red) or de-equalized (blue) HRTF sets for four different sampling schemes (color saturation). The test grid Ω_t was always the Lebedev$_{2702}$ grid. (a) Spectral differences ΔG across N over the full audio bandwidth. (b) Frequency-dependent spectral differences $\Delta G_f(\omega)$ at $N = 7$.

4.2 Binaural cues

Next, we compared the ILDs and ITDs of the reference HRTF set to those of order-limited or de-equalized sets, again with respect to different sampling schemes. For this, HRTFs in the horizontal plane ($\theta = 0°$) with an angular spacing of $\phi = 1°$ were extracted from the reference set $h_{REF,nm}$ and, depending on N, from the respective order-limited or de-equalized set $h_{OL,nm}$ and $h_{DEQ,nm}$. The broadband ILDs were then calculated as the ratio between the energy of the left and right ear HRIR (HRIR, the time-domain equivalent of an HRTF). The ITDs were calculated by means of a threshold-based onset detection on the ten times up-sampled and low-pass filtered HRIRs (10th order Butterworth low-pass at $3\,kHz$).

Figure 4 illustrates the calculated ILDs and ITDs of the reference HRTF set as well as of the order-limited and de-equalized sets, again exemplarily at $N = 7$. As can be seen in Fig. 4 (a), the ILDs of the de-equalized HRTFs are in good agreement with the reference and mostly unaffected by the sampling scheme. Overall, the Fliege grid shows the most notable deviations, especially at lateral directions. At these directions, also the Gaussian scheme provides clear deviations from the reference, whereas the equiangular and Lebedev schemes show only slight differences over the entire angular range. In contrast, the ILDs of the order-limited HRTFs (see Fig. 4 (b)) differ significantly from the reference. However, there is also only a rather weak influence of the sampling scheme. Regarding the ITDs, Fig. 4 (c) and (d) illustrate that there is virtually no influence of the sampling scheme, regardless of the upsampling method. Thus, at $N = 7$, the ITDs of the de-equalized HRTFs (see Fig. 4 (c)) as well as of the order-limited HRTFs (see Fig. 4 (d)) are in good agreement with the reference.

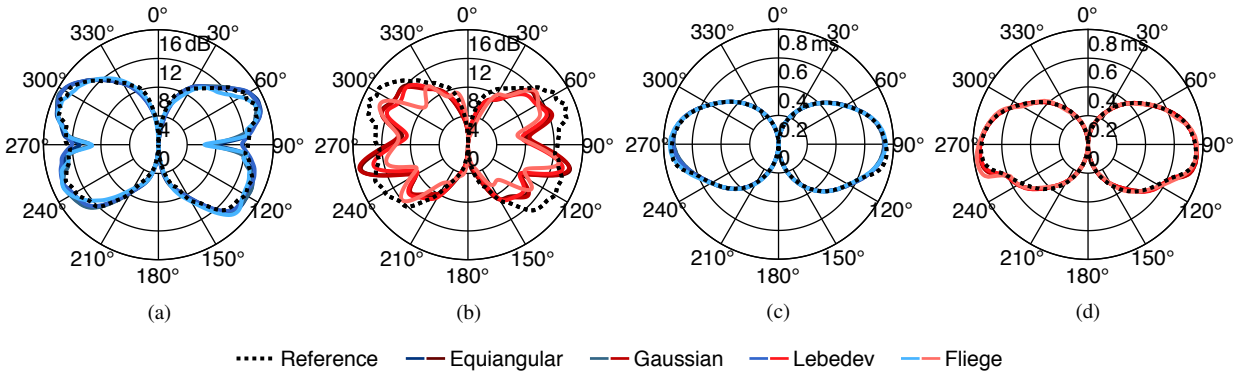


Figure 4. ILDs (a), (c) and ITDs (b), (d) in the horizontal plane of the reference (black) HRTFs as well as of the order-limited (red) or de-equalized (blue) HRTFs for four sampling schemes (color saturation) at $N = 7$.

4.3 Localization performance

To conclude the analysis, we compared the localization performance of order-limited and de-equalized HRTFs with respect to the sampling scheme applying two different auditory models from the Auditory Modeling Toolbox [14]. To assess the localization performance in the median sagittal plane, we used the model from Baumgartner et al. [2], which provides estimates for the polar RMS error (PE) as well as for the quadrant error rate (QE) based on monaural spectral cues. To evaluate the performance in the horizontal plane, we applied the model from May et al. [10], which estimates the azimuthal position of a sound source based on binaural cues. By comparing the intended and the estimated source position, a lateral error (LE) can be calculated. To calculate the error measures, first the performance of $h_{REF,nm}$, $h_{OL,nm}$, and $h_{DEQ,nm}$ was determined for each sampling scheme as a function of N. To estimate median plane localization performance, we used a test sampling grid Ω_t with $\phi = \{0°, 180°\}$ and $-30° \leq \theta \leq 90°$ in steps of 1°, and assumed a median listener sensitivity of $S = 0.76$. To estimate the horizontal plane localization performance, we applied a test sampling grid with $\phi = \pm 90°$ in steps of 5°. As final error measures, the absolute polar error difference (in degree)

$$\Delta PE = | PE_{REF} - PE_{TEST} |, \tag{8}$$

the absolute quadrant error difference (in percent)

$$\Delta QE = | QE_{REF} - QE_{TEST} |, \tag{9}$$

as well as the absolute lateral error difference (in degree)

$$\Delta LE = \frac{1}{T} \sum_{t=1}^{T} | LE_{REF}(\Omega_t) - LE_{TEST}(\Omega_t) |, \tag{10}$$

were calculated for each sampling scheme and order N, with the subscripts *REF* and *TEST* as defined above.

In the horizontal plane (see Fig. 5 (a)), the order-limited interpolation leads to an error increase at low orders $N \leq 4$. Similar to previous results, the Fliege grid performs worst here, especially at these low orders. In contrast, the SUpDEq method leads to hardly any increase in lateral error over the entire tested range of N, no matter which sampling scheme was applied. This shows that even at low orders, upsampling with the SUpDEq method always results in sufficient binaural cues.

In the median sagittal plane (see Fig. 5 (b) and (c)), order-limited interpolation leads to considerably higher errors over the entire range of N. Obviously, the high-frequency deviations in order-limited HRTFs badly affects the monaural spectral cues. Overall, the Fliege and the Lebedev grids seem to perform worse, but it is difficult to see a clear trend besides a general decrease in error with increasing order N. The SUpDEq method however amplifies the polar error only slightly at low orders $N \leq 2$. Once again, the Fliege grid tends to lead to a higher increase in error than the other grids, both for the polar error as well as for the quadrant error rate. The other grids perform more or less the same, with only slight increases in polar error and quadrant error rate at $N \geq 2$, indicating that spectral cues are only marginal impaired.

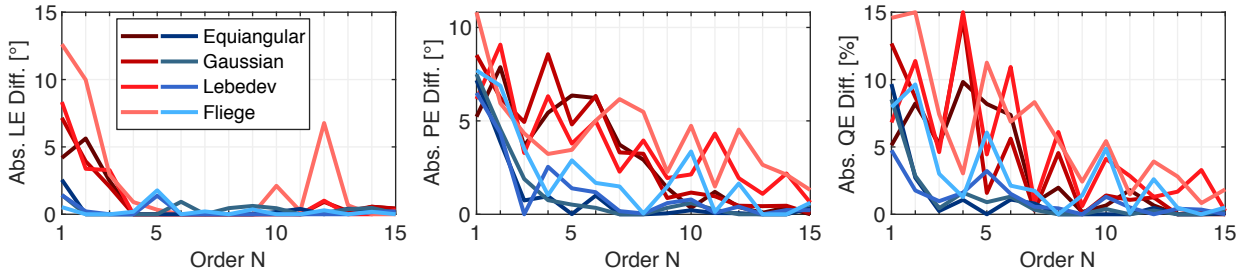


Figure 5. Absolute lateral error difference ΔLE (a), polar error difference ΔPE (b), and quadrant error difference ΔQE (c) across N for four different sampling schemes (color saturation) applying order-limited interpolation (red) or the SUpDEq method (blue).

5 CONCLUSION

This paper presented further evaluation of the SUpDEq method for spatial upsampling of sparse (individual) HRTF sets by investigating the influence of the spherical sampling scheme on the performance of the method. The study compared spatially upsampled HRTF sets originally based on sparse equiangular, Gaussian, Lebedev, and Fliege grids at various spatial orders N to a dense reference HRTF set, applying the SUpDEq method as well as common SH interpolation for upsampling. The analysis of spectral features, binaural cues, and localization performance revealed that the influence of the sampling scheme on the results of SUpDEq processing

is only marginal. Overall, only the Fliege scheme tended to perform a little worse than the other three tested schemes. With order-limited interpolation, the sampling scheme affected the examined features slightly stronger. The results of this study confirm or at least increase the general applicability of the SUpDEq method regarding the sampling scheme of the input HRTF set. Thus, the SUpDEq method might be applied with any sparse HRTF set measured on a proper full-spherical sampling grid. However, we only examined sampling schemes providing a closed-form expression in this paper. Therefore, further tests with arbitrary sampling schemes could be performed, even though it seems that given a reasonable sparse sampling scheme providing a well-conditioned (inverse) SH matrix, the results will be quite similar. Furthermore, listening experiments could be performed to analyze the perceptual influence of the sampling scheme, although the analysis in this paper suggests that the perceptual influence might be marginal. A Matlab-based implementation of the SUpDEq method is available on `https://github.com/AudioGroupCologne/SUpDEq`. The research presented in this paper was funded by the German Federal Ministry of Education and Research (BMBF 03FH014IX5-NarDasS).

REFERENCES

[1] V. R. Algazi, C. Avendano, and R. O. Duda. Estimation of a Spherical-Head Model from Anthropometry. *J. Audio Eng. Soc.*, 49(6):472–479, 2001.

[2] R. Baumgartner, P. Majdak, and B. Laback. Modeling sound-source localization in sagittal planes for human listeners. *J. Acoust. Soc. Am.*, 136(2):791–802, 2014.

[3] Z. Ben-Hur, D. L. Alon, B. Rafaely, and R. Mehra. Loudness stability of binaural sound with spherical harmonic representation of sparse head-related transfer functions. *EURASIP J. Audio, Speech, Music Process.*, 2019(5):1–14, 2019.

[4] B. Bernschütz. A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100. In *Proc. 39th DAGA*, pages 592–595, 2013.

[5] J. Blauert. *Spatial Hearing*. MIT Press, Cambridge, MA, 1996.

[6] R. Bomhardt, M. de la Fuente Klein, and J. Fels. A high-resolution head-related transfer function and three-dimensional ear model database. *Proc. Meet. Acoust.*, 29(1):1–11, 2017.

[7] F. Brinkmann and S. Weinzierl. Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition. In *Proc. Audio Eng. Soc. Conf. Audio for Virtual and Augmented Reality*, pages 1–10, 2018.

[8] J. Fliege and U. Maier. The distribution of points on the sphere and corresponding cubature formulae. *SIAM J. Numer. Anal.*, 19(2):317–334, 1999.

[9] V. I. Lebedev. Spherical quadrature formulas exact to orders 2529. *Siberian Math. J.*, 18(1):132–142, 1977.

[10] T. May, S. Van De Par, and A. Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(1):1–13, 2011.

[11] C. Pörschmann, J. M. Arend, and F. Brinkmann. Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling. *IEEE Trans. Audio, Speech, Lang. Process.*, 27(6):1060–1071, 2019.

[12] B. Rafaely. Analysis and Design of Spherical Microphone Arrays. *IEEE Trans. Speech, Audio Process.*, 13(1):135–143, 2005.

[13] B. Rafaely. *Fundamentals of Spherical Array Processing*. Springer-Verlag, Berlin Heidelberg, 2015.

[14] P. Søndergaard and P. Majdak. The Auditory Modeling Toolbox. In J. Blauert, editor, *The Technology of Binaural Listening*, pages 33–56. Springer-Verlag, Berlin Heidelberg, 2013.

2.4 Spatial Upsampling of Individual Sparse Head-Related Transfer Function Sets by Directional Equalization

Pörschmann, C., Arend, J. M., & Brinkmann, F. (2019). In *Proc. of the 23rd International Congress on Acoustics (ICA), Aachen, Germany* (pp. 4870–4877). https://doi.org/10.18154/RWTH-CONV-239484

PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Spatial upsampling of individual sparse head-related transfer function sets by directional equalization

Christoph PÖRSCHMANN[(1)†], Johannes M. AREND[(1)(2)], Fabian BRINKMANN[(2)]

[(1)]Institute of Communications Engineering, TH Köln, D-50679 Cologne, Germany,
[(2)]Audio Communication Group, TU Berlin, D-10587 Berlin, Germany
[†]Corresponding author, E-mail: christoph.poerschmann@th-koeln.de

Abstract

Determining full-spherical individual sets of head-related transfer functions (HRTFs) based on sparse measurements is a prerequisite for various applications in virtual acoustics. However, when applying HRTF interpolation in the spatially continuous spherical harmonics (SH) domain, the number of measured HRTFs limits the maximal accessible SH order. This results in a restricted spatial resolution and can cause perceptual artefacts like coloration or localization errors. In a previous publication we presented the SUpDEq method (Spatial Upsampling by Directional Equalization), which reduces these artifacts by a directional equalization based on a spherical head model prior to the SH transform. This removes direction-dependent temporal and spectral components and thus reduces the spatial complexity of the HRTF set enabling improved interpolation of HRTFs already at low SH orders. A subsequent de-equalization recovers energy in higher spatial orders that was discarded in the sparse HRTF set. In this study we analyze 96 individual HRTF sets and investigate to what extent the performance of SUpDEq, which we already analyzed for dummy heads, can be transferred to individual HRTF sets. The results show that the SUpDEq method clearly outperforms common SH interpolation of individual HRTFs with respect to the spectral structure and to modeled localization performance.

Keywords: Binaural hearing, Localization, Head-related transfer functions, Virtual acoustic environments

1 INTRODUCTION

A spatial presentation of sound sources is a fundamental element of virtual acoustic environments (VAEs). For this, monaural and binaural cues, which are mainly caused by the shape of the pinna and the head, need to be considered. While spectral information serves as main cue to determine elevation, differences between the signals reaching the left and the right ear allow lateral localization. These binaural cues manifest in interaural time differences (ITDs) and interaural level differences (ILDs). In many headphone-based VAEs, head-related transfer functions (HRTFs) are applied to describe the sound incidence from a source, which is typically in the far-field, to the left and right ear incorporating both, monaural and the binaural cues. Generally, the use of individual HRTFs is advantageous, for example regarding localization accuracy in the median plane [5]. However, a high number of HRTFs is required to adequately capture the relevant cues for all directions of incidence which makes the measurements time-consuming and tedious.

To allow an optimized interpolation between the measured directions, complete sets of HRTFs can be measured on a spherical grid and described in the spherical harmonics (SH) domain [14, 12]. In this case a decomposition into spherical base functions of different spatial orders N is applied, where higher orders correspond to a higher spatial resolution. A subsequent inverse spatial Fourier transform at arbitrary angles can be used to recover a spatially upsampled HRTF set. However, describing sparse HRTF sets in the SH domain results in a limited spatial order and incorporates an incomplete description of the spatial properties resulting in spatial aliasing or truncation errors. To avoid spatial alisaing, an order $N \geq kr$ with $k = \omega/c$, and r being the head radius is required [11, 4]. For the full audio bandwidth ($f \leq 20\,kHz$) this leads to $N = 32$ requiring at least 1089 measured directions when assuming $r = 8.75$ cm and $c = 343$ m/s.

Different studies analyzed the artifacts of sparsely measured HRTF sets or examined methods to reduce them

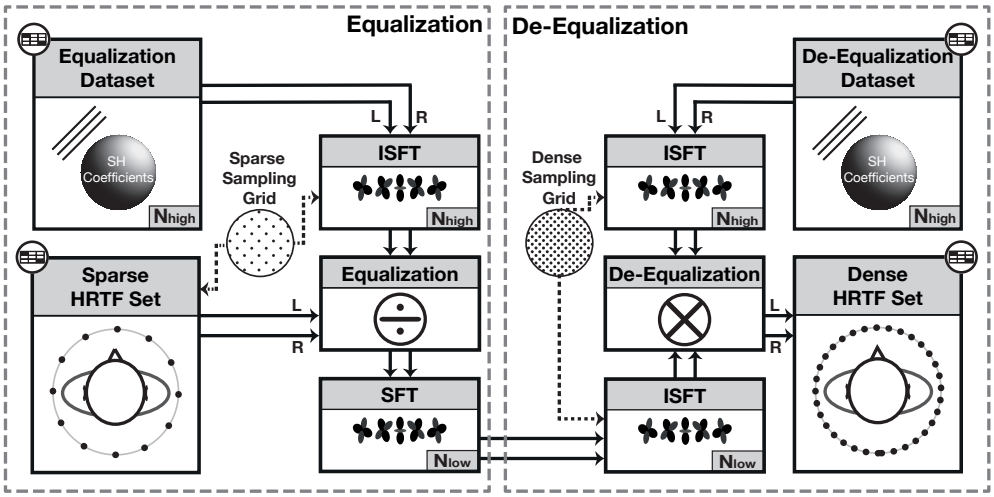


Figure 1. Block diagram of the SUpDEq method. Left panel: A sparse HRTF set is equalized on the corresponding sparse sampling grid before transformed to the SH domain with $N = N_{low}$. Right panel: The equalized set is de-equalized on a dense sampling grid. If required, the resulting dense HRTF set can again be transformed to the SH domain with $N = N_{high}$.

(e.g. [4, 3, 15, 7]). In this scope we recently introduced the SUpDEq (Spatial Upsampling by Directional Equalization) method [10], which removes frequency-dependent ITDs and ILDs as well as head-related elevation-dependent spectral features from the HRTFs. SUpDEq applies a spectral division (equalization) of the HRTF with a corresponding equalization function prior to the SH transform. A directional rigid sphere transfer function can be used here as equalization function, resulting in a significantly reduced spatial order N. After spatial upsampling, a de-equalization by means of a spectral multiplication with the same equalization function recovers a spatially upsampled HRTF set. In this paper we analyze the SUpDEq method for a large number of measured and simulated datasets.

2 METHOD

The SUpDEq method has been described in detail in [10]. In the following we thus briefly outline the basic concept. The corresponding block diagram is given in Fig. 1. First, the sparse HRTF set H_{HRTF} measured at S sampling points $\Omega_s = \{(\phi_1, \theta_1), \ldots, (\phi_S, \theta_S)\}$ is spatially equalized with an appropriate equalization dataset H_{EQ}

$$H_{HRTF,EQ}(\omega, \Omega_s) = \frac{H_{HRTF}(\omega, \Omega_s)}{H_{EQ}(\omega, \Omega_s)}. \tag{1}$$

While generally different equalization datasets can be applied, in this study a rigid sphere transfer function is used [14, p. 227]. The radius of the sphere corresponds to the physical dimensions of a human head, as ear position $\phi = \pm 90^\circ$ and $\theta = 0^\circ$ is considered. The rigid sphere transfer function can thus be regarded as a simplified HRTF set featuring basic temporal and spectral components, but leaving out information on the shape of the outer ears or the fine structure of the head. Thus, by the equalization a time-alignment of the HRTFs is performed and direction-dependent influences of the spherical shape of the head are compensated. As a consequence, the equalization with the rigid sphere transfer function considerably reduces the directional complexity of $H_{HRTF,EQ}$ and thus the required order for the SH transform. As the equalization dataset H_{EQ} can be calculated based on an analytical description, it can be determined at a freely chosen maximal order, typically $N_{high} \geq 35$. The SH coefficients for the equalized sparse HRTF set are obtained by applying the SH transform

on the equalized HRTFs up to an appropriate low maximal order N_{low} which corresponds to the maximal order that can be resolved by Ω_s. Then an upsampled HRTF set $\widehat{H}_{HRTF,EQ}$ is calculated on a dense sampling grid $\Omega_d = \{(\phi_1, \theta_1), \ldots, (\phi_D, \theta_D)\}$, with $D \gg S$ by using the inverse SH transform. Finally, HRTFs are reconstructed by a subsequent de-equalization by means of spectral multiplication with a de-equalization dataset H_{DEQ}

$$\widehat{H}_{HRTF,DEQ}(\omega, \Omega_d) = \widehat{H}_{HRTF,EQ}(\omega, \Omega_d) \cdot H_{DEQ}(\omega, \Omega_d). \tag{2}$$

For de-equalization, again the rigid sphere transfer function is used in the present study. This last step recovers energy at higher spatial orders that was transformed to lower orders within the equalization. Again, $H_{HRTF} = \widehat{H}_{HRTF,DEQ}$ holds if N_{low} and N_{high} are chosen appropriately. Energy which, after the equalization, still is apparent at high modal orders $N > N_{low}$ results in spatial aliasing and truncation errors as it is irreversibly mirrored to lower orders $N \leq N_{low}$ [4]. Thus we obtain $H_{HRTF} \approx \widehat{H}_{HRTF,DEQ}$. The following section analyzes the influence of these deviations for individual datasets and investigates which advantage the SUpDEq method provides compared to common (order-limited) SH interpolation without any pre- or postprocessing.

3 EVALUATION

In previous publications [10, 9] we investigated the performance of SUpDEq for different artificial heads. However, one of the target applications of the SUpDEq method is the reduction of the measurement effort of individual HRTF sets. Thus, in this study we analyze the performance of the SUpDEq method for the HUTUBS database which is online available on `http://dx.doi.org/10.14279/depositonce-8487`. The database contains of 96 acoustically measured and 96 numerically simulated datasets of full-spherical HRTFs (94 subjects plus 2 repeated measurements of a human subject and an artificial head). For more detailed information on the database please refer to [6]. We apply the HRTF sets to compare the performance of the SUpDEq method (de-equalized HRTFs) to HRTFs obtained with with strictly order limited SH interpolation, i.e., without any pre- or post-processing before or after the SH transform. For this we generated SH coefficients from 15 sparse sampling grids equaling (limited) orders of $N = 1 - 15$. Thus, both order-limited (OL) and de-equalized (DEQ) sets are based on the same respective sparse grid. To generate various sparse HRTF sets which we used as input data for the evaluation, we simply spatially subsampled each individual reference set in the SH domain by means of the inverse SH transform at the required directions. We calculated the optimal radius for the rigid sphere model for each of the sets according to Algazi et al. [1] based on the individual anthropometry resulting in an average value over the complete set of $r = 9.1\,\text{cm}$ ($SD = 0.23\,\text{cm}$).

3.1 Spectral differences

First we analyze the spectral deviations to the reference set as a function of N on various test sampling grids with T sampling points $\Omega_t = \{(\phi_1, \theta_1), \ldots, (\phi_T, \theta_T)\}$. For this the frequency-dependent spectral differences per sampling point were calculated in dB as

$$\Delta g(\omega, \Omega_t) = 20lg \frac{|\, H_{HRTF,REF}(\omega, \Omega_t)\,|}{|\, H_{HRTF,TEST}(\omega, \Omega_t)\,|}, \tag{3}$$

where $H_{HRTF,REF}$ is the left ear HRTF extracted from the reference set and $H_{HRTF,TEST}$ the one extracted from the order-limited or the de-equalized datasets at the sampling point Ω_t. Then, the absolute value of $\Delta g(\omega, \Omega_t)$ was averaged across the temporal frequency ω to obtain one value $\Delta G_{sp}(\Omega_t)$ (in dB) per sampling point

$$\Delta G_{sp}(\Omega_t) = \frac{1}{n_\omega} \sum_{\omega=1}^{n_\omega} |\, \Delta g(\omega, \Omega_t)\,|, \tag{4}$$

across all sampling points Ω_t to obtain the frequency-dependent measure $\Delta G_f(\omega)$ (in dB)

$$\Delta G_f(\omega) = \frac{1}{n_{\Omega_t}} \sum_{\Omega_t=1}^{n_{\Omega_t}} |\, \Delta g(\omega, \Omega_t)\,|, \tag{5}$$

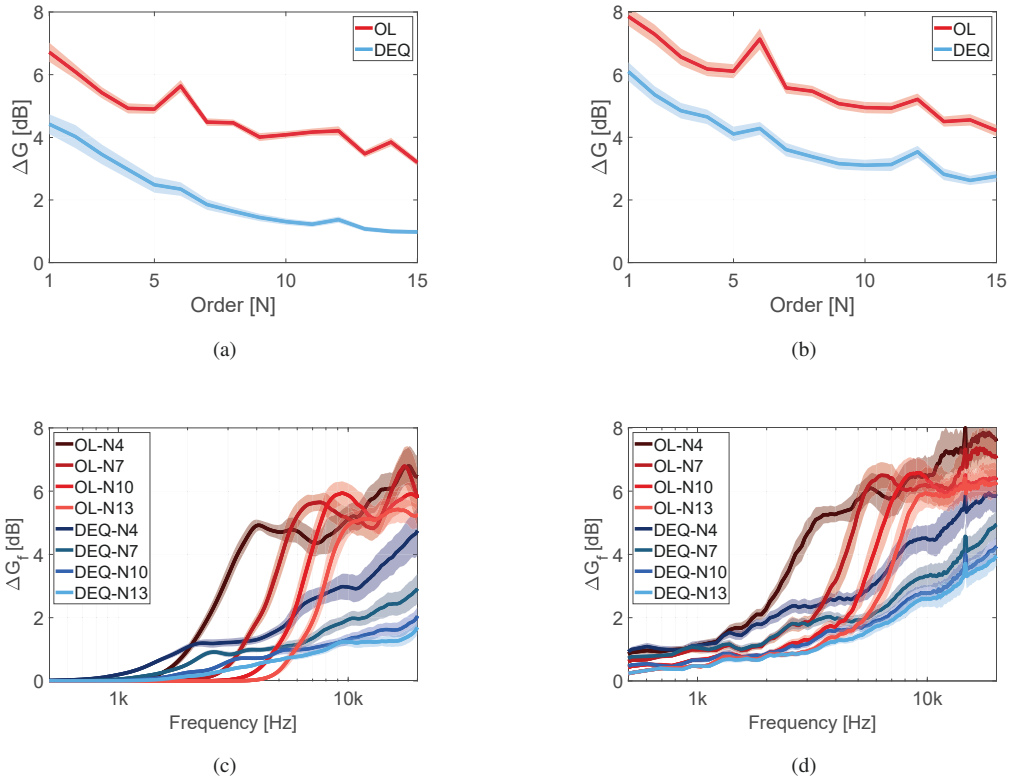


Figure 2. Spectral differences in dB (left ear) between the reference HRTF sets and the order-limited (OL) or de-equalized HRTF sets (DEQ), both based on the respective sparse set, averaged over all 96 datasets. Additionally the standard deviations are plotted (shaded). The left row (a,c) illustrates the results for the simulated datasets, in (b,d) the ones for the measured datasets are given. In (a,b) the spectral differences ΔG averaged over the full audio bandwidth across N for order-limited datasets (red) and the de-equalized datasets (blue) are given, in (c,d) the frequency-dependent spectral differences $\Delta G_f(\omega)$ for $N = 4, 7, 10, 13$ (color saturation).

and across ω and Ω_t, resulting in a single value ΔG (in dB) describing the spectral difference

$$\Delta G = \frac{1}{n_{\Omega_t}} \frac{1}{n_\omega} \sum_{\Omega_t=1}^{n_{\Omega_t}} \sum_{\omega=1}^{n_\omega} | \Delta g(\omega, \Omega_t) | . \tag{6}$$

Finally, the average values and standard deviations over all 96 datasets were calculated for the simulated and the measured datasets.

Fig. 2 (a,b) show the spectral differences ΔG across N for order-limited interpolation and the SUpDEq method (de-equalized datasets) over the full audio bandwidth using the reference Lebedev$_{2702}$ grid as test sampling grid Ω_t. The SUpDEq method clearly outperforms the order-limited interpolation both for the simulated and the mesaured HRTF sets. The spectral differences are about $2-3$ dB lower than for order-limited interpolation. Fig. 2 (c,d) show the frequency-dependent spectral differences $\Delta G_f(\omega)$ at $N = 4, 7, 10, 13$. Generally, the spectral differences are quite small at low frequencies. For order-limited interpolation they suddenly rise within one octave from about 2 dB up to about 5 dB or more above a specific alias frequency. For the SUpDEq method, however, the spectral differences show a much more gentle rise. The differences exceed 2 dB for frequencies

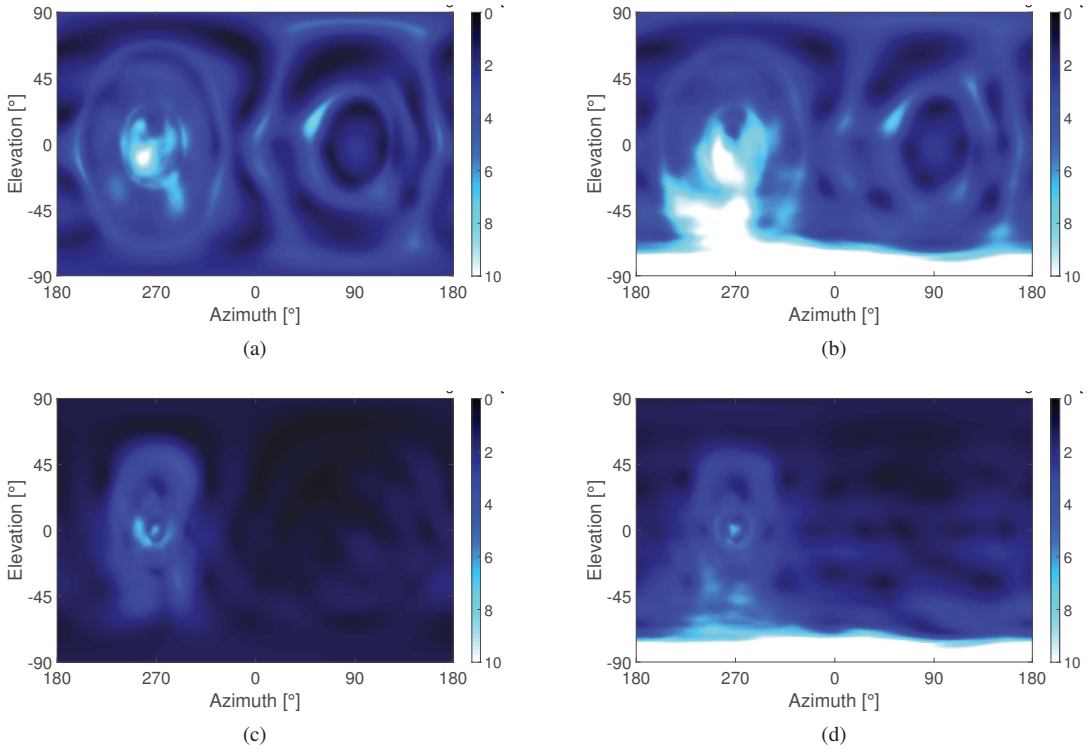


Figure 3. Spectral differences $\Delta G_{sp}(\Omega_t)$ per sampling point for order-limited interpolation (a,b) and for the SUpDEq method (c,d) at $N = 4$ and $f \leq 10\,\text{kHz}$ averaged over all 96 datasets. The left row (a,c) shows the results for the simulated datasets, the right row (b,d) the results for the measured ones.

above 3 kHz for $N = 4$, while differences stay below 2 dB for orders of $N \geq 10$ up to 10 kHz (DEQ).

Fig. 3 concludes the spectral analysis and shows the spectral differences $\Delta G_{sp}(\Omega_t)$ per sampling point at $N = 4$, $f \leq 10\,\text{kHz}$, and a full spherical test sampling grid Ω_t with a resolution of 1° in azimuth and elevation. As depicted in Fig. 3 (a,b), the order-limited interpolation results in distinct spectral differences spread over the entire angular range. On the contrary, Fig. 3 (c,d) shows that for the SUpDEq method the spectral differences are mainly located at contralateral directions. At frontal directions, where order-limited interpolation typically performs badly, the SUpDEq method shows good results. The same can be observed for various ipsilateral directions. The spectral differences are generally higher for order-limited interpolation, with a maximum of about $\Delta G_{sp}(\Omega_t) = 10.4\,\text{dB}$ at $\phi = 262^\circ$ and $\theta = -10^\circ$ averaged over all subjects for the simulated datasets. For these datasets applying the SUpDEq method results in a maximal spectral difference $\Delta G_{sp}(\Omega_t)$ of 6.6 dB at $\phi = 257^\circ$ and $\theta = 2^\circ$. Finally, Fig. 3 (b,d) show the same trend for the measured datasets, but reveal large deviations for the downward directions. This is caused by the acoustic shadowing of the measurement equipment and is described in detail in [6].

3.2 Localization performance

To compare the localization performance of order-limited HRTFs and de-equalized HRTFs in the median sagittal plane, we used the model from Baumgartner et al. [2] which compares the spectral structure of a reference HRTF set to a set of test HRTFs. Based on a probabilistic estimate of the perceived sound source location, the model determines the polar RMS error which describes the expected angular error between the actual and

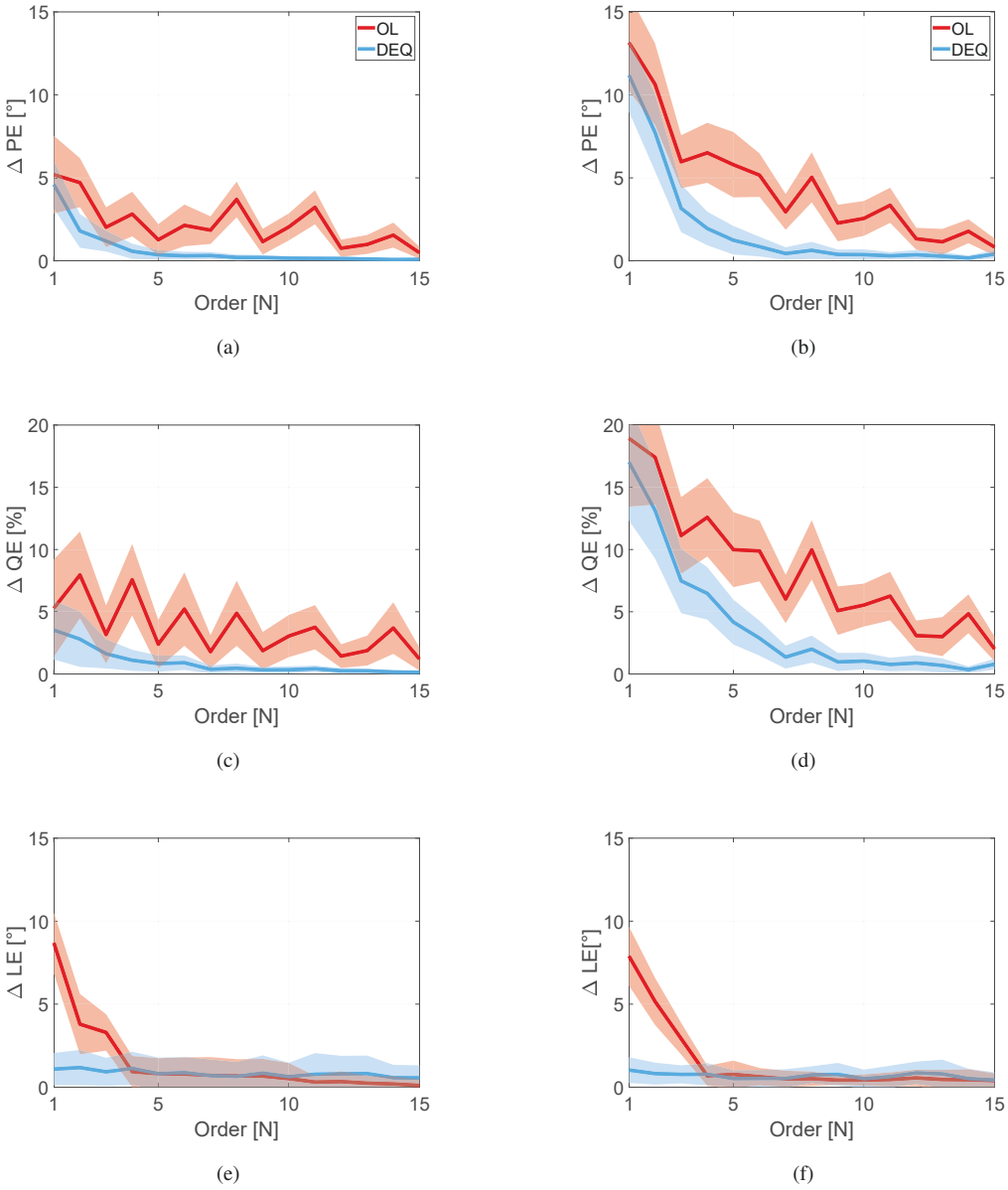


Figure 4. Absolute polar error difference ΔPE (a,b), quadrant error difference ΔQE (c,d), and lateral error difference ΔLE (e,f) over SH order N for order-limited interpolation (red) and the SUpDEq method (blue) averaged over the 96 individual datasets. Additionally, the standard deviations are shown (shaded). In the left row (a,c,e) the results for the simulated HRTF sets are shown, in the right row (b,d,f) the results for the measured HRTF sets.

perceived source positions. Additionally, it determines the quadrant error rate which specifies the front-back and up-down confusions. Regarding the localization performance in the horizontal plane, we used the model from May et al. [8] which weighs the frequency-dependent binaural cues (ILDs, ITDs) to estimate the azimuthal

position of a sound source. A lateral error can be calculated by comparing the intended and the estimated source position. For the analysis of both models we used the Auditory Modeling Toolbox (AMT) [13]. The procedure for determining the errors has been described in detail in [10] and can be outlined as follows. To estimate median sagittal plane localization performance, we used a test sampling grid Ω_t with $\phi = \{0°, 180°\}$ and $-30° \leq \theta \leq 90°$ in steps of 1°, and assumed a median listener sensitivity of $S = 0.76$ (according to Baumgartner et al. [2]). For the horizontal plane localization performance, we used a test sampling grid with $\phi = \pm 90°$ in steps of 5°. We determined the absolute polar error difference (PE in degree)

$$\Delta PE = \mid PE_{REF} - PE_{TEST} \mid, \tag{7}$$

the absolute quadrant error difference (QE in percent)

$$\Delta QE = \mid QE_{REF} - QE_{TEST} \mid, \tag{8}$$

as well as the absolute lateral error difference (LE in degree)

$$\Delta LE = \frac{1}{T} \sum_{t=1}^{T} \mid LE_{REF}(\Omega_t) - LE_{TEST}(\Omega_t) \mid, \tag{9}$$

for each order N with the subscripts *REF* describing the referene dataset and *TEST* the dataset under test. Again we calculated the averages and standard deviations over all datasets separated for the simulated and the measured sets.

As plotted in Fig. 4 (a–d), in the median sagittal plane the order-limited interpolation leads both for the simulated and the measured datasets to higher errors than the SUpDEq method. High-frequency deviations of the order-limited HRTFs affect spectral cues which are relevant for sagittal plane localization. For the de-equalized datasets, ΔPE decreases with increasing order N, $\Delta PE \leq 2°$ holds for $N \geq 4$. Thus the spectral cues seem to be mostly unimpaired here. The extent of the quadrant error ΔQE varies greatly between the measured and the simulated sets and lies for the order-limited sets between 4 % (simulated) and 10 % (measured) at $N \geq 7$. However, for the de-equalized datasets, ΔQE is below 2 % at $N \geq 7$. Generally, in the median sagittal plane the average errors are much higher for the measured datasets than for the simulated ones. This is probably a result of the measurement inaccuracies for downward directions, which as well have been observed in Sec. 3. In Fig. 4 (e–f) the localization performance in the horizontal plane is shown. Here the order-limited interpolation performs quite well, even though lateral errors are distinctly amplified at orders $N \leq 3$. This might be caused by strong pre-ringing artifacts causing wrong ITDs, as already discussed in [10]. The SUpDEq method leads to hardly any increase in lateral error over the entire tested range of N.

4 CONCLUSION

In this paper we analyzed the performance of the SUpDEq method for spatial upsampling of individual sparse HRTF sets. Regarding the spectral structure, the deviations from the reference HRTF set are significantly smaller for the SUpDEq method than for order-limited interpolation. The average difference is about 2 dB, both for the simulated and the measured datasets. Furthermore, the analysis of the spectral diffences showed for the SUpDEq methods a much more gentle rise over frequency than for the order-limited interpolation. Finally, the spectral differences induced by the SUpDEq method are mainly at contralateral directions, while the differences due to order-limited interpolation spread over the entire angular range, with distinct clusters at frontal and contralateral directions. Regarding the modeled localization performance the SUpDEq method performed better in both planes because spectral and binaural cues are less impaired in comparison to the order-limited interpolation.

Generally, the evaluation showed that the results found for dummy heads in [10] can be generalized to individually measured or simulated datasets. Thus, the SUpDEq approach can help closing the gap between a practical and fast measurement procedure and sufficient accuracy of the upsampled HRTF set. However, for such a simplified procedure other influencing factors like e.g. the elimination of room reflections [9] or the compensation of small displacements of the human head during the measurement need to be considered.

The research presented in this paper has been funded by the German Federal Ministry of Education and Research. Support Code: BMBF 03FH014IX5-NarDasS. A Matlab-based implementation of the SUpDEq method is available on `https://github.com/AudioGroupCologne/SUpDEq`.

REFERENCES

[1] V. Algazi, C. Avendano, and R. O. Duda. Estimation of a Spherical-Head Model from Anthropometry. *J. Audio Eng. Soc.*, 49(6):472 – 479, 2001.

[2] R. Baumgartner, P. Majdak, and B. Laback. Modeling sound-source localization in sagittal planes for human listeners. *J. Acous. Soc. Am.*, 136(2):791–802, 2014.

[3] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely. Spectral equalization in binaural signals represented by order-truncated spherical harmonics. *The Journal of the Acoustical Society of America*, 141(6):4087–4096, 2017.

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

[5] J. Blauert. *Spatial Hearing - The Psychophysics of Human Sound Localization*. MIT Press, Cambridge, MA, revised edition, 1996.

[6] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl. A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses. *Journal of the Audio Engineering Society, in press*, 2019.

[7] F. Brinkmann and S. Weinzierl. Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition. In *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality*, pages 1–10, 2018.

[8] T. May, S. Van De Par, and A. Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(1):1–13, 2011.

[9] C. Pörschmann and J. M. Arend. Obtaining Dense HRTF Sets from Sparse Measurements in Reverberant Environments. In *Proceedings of the AES Conference on Immersive and Interactive Audio*, 2019.

[10] C. Pörschmann, J. M. Arend, and F. Brinkmann. Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 27(6):1060 – 1071, 2019.

[11] B. Rafaely. Analysis and Design of Spherical Microphone Arrays. *IEEE Transaction on Speech and Audio Processing*, 13(1):135–143, 2005.

[12] B. Rafaely. *Fundamentals of Spherical Array Processing*. Springer-Verlag, Berlin Heidelberg, 2015.

[13] P. Søndergaard and P. Majdak. The Auditory Modeling Toolbox. In J. Blauert, editor, *The Technology of Binaural Listening*, pages 33–56. Springer-Verlag, Berlin Heidelberg, 2013.

[14] E. G. Williams. *Fourier Acoustics - Sound Radiation and Nearfield Acoustical Holography*. Academic Press, London, UK, 1999.

[15] M. Zaunschirm, C. Schoerkhuber, and R. Hoeldrich. Binaural rendering of Ambisonic signals by HRIR time alignment and a diffuseness constraint. *J. Acous. Soc. Am.*, 143(6):3616 – 3627, 2018.

2.5 Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling

Pörschmann*, C., Arend*, J. M., & Brinkmann, F. (2019). *IEEE/ACM Trans. Audio Speech Lang. Proc.*, *27*(6), 1060–1071. (*equal contributions). https://doi.org/10.1109/TASLP.2019.2908057

Pörschmann*, C., Arend*, J. M., & Brinkmann, F. (2020). Correction to "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling." *IEEE/ACM Trans. Audio Speech Lang. Proc.*, *28*, 2194–2194. (*equal contributions). https://doi.org/10.1109/TASLP.2020.3010608

Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling

Christoph Pörschmann, Johannes M. Arend, and Fabian Brinkmann

***Abstract*—Acquiring decent full-spherical sets of head-related transfer functions (HRTFs) based on a small number of measurements is highly desirable. For spatial upsampling, HRTF interpolation in the spatially continuous spherical harmonics (SH) domain is a common approach. However, the number of measured HRTFs limits the assessable SH order, resulting in order-limited HRTFs when transformed to the SH domain. Thus, the SH representation of sparse HRTF sets shows restricted spatial resolution and suffers from order-limitation errors. We present a method that reduces these errors by a directional equalization prior to the SH transform. This is done by a spectral division of each HRTF with a corresponding directional rigid sphere transfer function. The processing removes direction-dependent temporal and spectral components and, therefore, significantly reduces the spatial complexity of the HRTF set, allowing for an enhanced interpolation of HRTFs at reduced SH orders. Spatial upsampling is achieved by an inverse SH transform on an arbitrary dense sampling grid. A subsequent de-equalization by a spectral multiplication with the rigid sphere transfer function recovers the energy in higher spatial orders that was not inherent in the sparse HRTF set. For evaluation, HRTFs were calculated for various limited orders from sparse datasets and compared to a reference. The results show that the proposed method clearly outperforms common SH interpolation of HRTF spectra regarding the overall spectral and temporal structure as well as modeled localization performance.**

***Index Terms*—Spatial audio, spherical harmonics, head-related transfer functions (HRTFs).**

I. Introduction

LOCALIZATION of sound is one of the outstanding skills of the human auditory system. For this, humans use monaural and binaural cues, which are mainly caused by the shape of the pinna, the head, and to some extent by the torso. Whereas spectral information is the main cue to determine elevation (monaural cues), humans use differences between the signals reaching the left and the right ear for lateral localization (binaural cues). These differences manifest in interaural time differences (ITDs) and interaural level differences (ILDs). Head-related transfer functions (HRTFs), which describe the sound incidence from a sound source (typically in the far-field) to the left and right ear, incorporate both, monaural and binaural cues, and are essential when realizing headphone-based virtual acoustic environments [1]–[3].

To adequately present spatial cues for all directions of incidence, an appropriately large number of HRTFs is necessary due to the high sensitivity of the human auditory system. Such a set of HRTFs describes the sound reaching a listener from various directions and can, for example, be measured on a spherical (spatial) sampling grid. Quite common spherical sampling grids for capturing a sound field are the so-called Lebedev or Gaussian grids [4]. Often, dummy heads are used to obtain measured HRTF sets [5]–[7], however, listening to nonindividual HRTFs for instance causes increased localization errors [8]. Individual HRTFs have, thus, to be used for high-fidelity spatial audio, which can be measured with specialized experimental setups [9]–[11].

Lindau and Weinzierl [12] found that a resolution of 2° in the horizontal and 1° in the vertical direction is required for an artifact-free dynamic auralization with HRTFs when no further interpolation of the HRTFs is applied, which results in an equal angle grid of about 32 000 HRTFs. This impractically high number, which is particularly adverse for individual measurements, can be reduced by applying physically and perceptually suitable interpolation methods. In other words, a so-called dense grid of HRTFs can be gained by measuring HRTFs for a limited number of positions on a so-called sparse grid, followed by spatial upsampling by means of interpolation. By this, the data size of measured sets can be decreased and, even more important, the interpolation approach can considerably reduce the effort to acquire (appropriate) dense HRTF sets.

Early attempts of such interpolation approaches either took place in the frequency or in the time domain. Wenzel and Foster [13] investigated the influence of interpolation on localization for nonindividual HRTFs. They found that even for large interpolation intervals, localization accuracy was largely unaffected. Chen *et al.* [14] investigated if HRTFs can be synthesized by feature extraction. They showed that a weighted combination

Manuscript received August 23, 2018; revised December 23, 2018 and February 25, 2019; accepted March 18, 2019. Date of publication March 28, 2019; date of current version April 24, 2019. This work was supported in part by the German Federal Ministry of Education and Research (BMBF 03FH014IX5-NarDasS) and in part by the German Research Foundation (DFG WE 4057/3-2). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Federico Fontana. *(Christoph Pörschmann and Johannes M. Arend contributed equally to this work.) (Corresponding author: Christoph Pörschmann.)*

C. Pörschmann is with the Institute of Communications Engineering, TH Köln - University of Applied Sciences, Cologne 50679, Germany (e-mail: christoph.poerschmann@th-koeln.de).

J. M. Arend is with the Institute of Communications Engineering, TH Köln - University of Applied Sciences, Cologne 50679, Germany, and also with the Audio Communication Group, Technical University of Berlin, Berlin 10587, Germany (e-mail: johannes.arend@th-koeln.de).

F. Brinkmann is with the Audio Communication Group, Technical University of Berlin, Berlin 10587, Germany (e-mail: fabian.brinkmann@tu-berlin.de).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. The material includes further evaluations of the presented method. Contact christoph.poerschmann@th-koeln.de for further questions about this paper.

Digital Object Identifier 10.1109/TASLP.2019.2908057

of eigen transfer functions can be used for HRTF interpolation. Langendijk and Bronkhorst [15] performed a study analyzing the perceptual influence of the resolution of measured HRTFs. Especially the initial time delay, which strongly varies depending on the angle of sound incidence, caused severe problems when interpolating in the time domain. Thus, various investigations have been performed in order to improve the interpolation of HRTFs: Hartung *et al.* [16] and Djelani *et al.* [17] described a method that removes the initial time delay separately for each channel of a head-related impulse response (HRIR, the time-domain equivalent of an HRTF) by determining the position of the absolute maximum. The interpolation of the spatially closest and time-aligned impulse responses was performed linearly and finally, an interpolated initial delay was added. Following this idea, Minnaar *et al.* [18] determined that a number of 1130 measured HRTFs is sufficient for an adequate representation in virtual acoustic environments. Hartung *et al.* [16] found that a separate spline-based interpolation of the magnitude and unwrapped phase response of the HRTF performed best, both in a technical and perceptual analysis. However, a direct comparison of the different interpolation approaches is very challenging as the accuracy strongly depends on the resolution and on the grid type.

Several authors suggested to describe HRTF sets in the spherical harmonics (SH) domain [19, ch. 6], [20, ch. 1]. Here, the HRTF set, measured on a spherical grid, is decomposed into spherical base functions of different orders N, where higher orders allow a higher spatial resolution. In this case, the number of measured directions directly corresponds to the maximum usable order N (see (4) in Section II). To entirely consider the spatial dependency of the HRTF set for the full audio bandwidth, a SH order $N \geq kr$ with $k = \frac{\omega}{c}$, ω the temporal frequency, c the speed of sound, and r the head radius is required [4], [21], [22]. Assuming $r = 8.75$ cm as the average human head radius [23] and $c = 343$ m/s leads to a minimum required spatial order $N = 32$ for performing a nearly perfect interpolation of HRTFs for frequencies up to 20 kHz. Depending on the spatial sampling grid, this would require at least 1089 measurements. Consequently, sparse HRTF sets result in a limited order in the SH domain. Such an order limitation implies an incomplete description of the spatial properties of the HRTF set and leads to spatial-aliasing artifacts such as impairments of high-frequency components and binaural cues [21], [24], [25].

Because the analytical SH basis functions are spatially continuous and solutions of the wave equation, the interpolation approach in the SH domain yields a physically correct and spatially continuous HRTF representation for $N \geq kr$. Thus, HRTFs can be obtained for arbitrary source positions by evaluating the SH functions at the corresponding directions. Already in 1998, Evans *et al.* [26] proposed an interpolation method using spherical harmonics either in time or frequency domain. Moreover, the authors investigated how eliminating ITDs could further enhance their approach. Since then, different approaches have been carried out in order to eliminate the perceptual artifacts of order-truncated and order-limited HRTF representations, with the aim to reduce the required measurement positions. Bernschütz *et al.* [21] proposed a spatial resampling of the HRTFs at the desired low order. By this, high-order HRTF energy is mirrored to lower orders, and the low-pass effect of order truncation is diminished. Ben-Hur *et al.* [25] suggested a post-hoc equalization for order-truncated HRTFs, which reduces coloration artifacts on average and shifts the largest errors from frontal to lateral source positions. Both approaches considerably decrease perceptual artifacts if using spatial SH orders of 6 or 7 – corresponding to at least 49 measurements. However, with both methods, the strength of the observed artifacts highly depends on the source position, at least for low SH orders.

Recent studies examined how a separate transform of magnitude and phase spectra [27], [28] and time aligning the HRTFs (by eliminating their linear phase components) before SH transform [22], [24], [28] affects the required SH order as well as auditory perception. The results suggest that these approaches reduce order-limitation artifacts and, moreover, considerably weaken the dependency on the source position. In addition, any approach that uses time alignment or considers a separate treatment of the magnitude information could be combined with a spatially continuous artificial phase derived either from analyzing HRIR onsets, or from anthropometric head measures [29]–[31]. This would also allow a post-hoc individualization of the ITD. However, other features, such as the fine structure of the magnitude and phase response, were not taken into account in the studies mentioned above, even though these features could also significantly affect the required SH order. In other words, instead of only eliminating the linear phase components of HRTFs, other typical HRTF features, which for example can be analytically described, could be additionally removed from the HRTFs. A SH transform of such datasets certainly results in a further reduced maximal SH order.

This is the principle of the spatial upsampling by directional equalization (SUpDEq) method presented in this paper. The method attempts to remove direction-dependent temporal and spectral components from the HRTFs before SH transform, such as frequency-dependent ITDs and ILDs as well as elevation-dependent spectral features. This is done by a spectral division (equalization) of the HRTFs with corresponding rigid sphere transfer functions (STFs), which can be regarded as a simplified HRTF set only comprising basic temporal and spectral features. From an information theory point of view, this procedure is equivalent to coding the prediction error between the HRTFs and the STFs, which typically has a much lower SH order than the original HRTF set. After spatial upsampling (SH interpolation), a de-equalization by means of a spectral multiplication with the same STFs is performed to recover a spatially upsampled HRTF set. In general, the proposed method of directional equalization of HRTF sets can be applied in combination with all spatial upsampling methods. However, as SH-based interpolation techniques can be regarded as state of the art [28], this paper focuses on spatial upsampling in the SH domain.

This paper is organized as follows. The proposed SUpDEq method is described in greater detail in Section II, followed by an evaluation of the approach in Section III. Section IV discusses the results of this evaluation. Finally, in Section V, we draw a conclusion and give an outlook on which applications can benefit from the presented method.

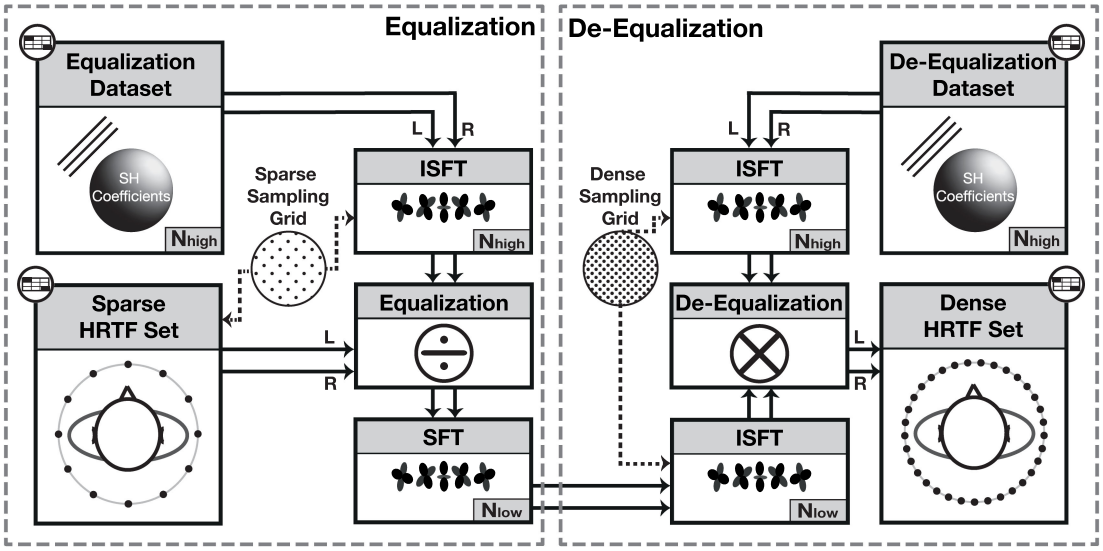


Fig. 1. Block diagram of the SUpDEq method. Left panel: A sparse HRTF set is equalized on the corresponding sparse sampling grid. The equalized set is then transformed to the SH domain with $N = N_{\text{low}}$. Right panel: The equalized set is de-equalized on a dense sampling grid, resulting in a dense HRTF set. The final dense HRTF set could then again be transformed to the SH domain with $N = N_{\text{high}}$.

II. Method

A spherical dataset $H(\omega, \Omega_g)$ can be described in the SH domain by the SH coefficients $f_{\text{nm}}(\omega)$ that are obtained via the SH transform, often also referred to as spatial (or spherical) Fourier transform [19, p. 2], [20, p. 16]

$$f_{\text{nm}}(\omega) = \sum_{g=1}^{G} H(\omega, \Omega_g) Y_n^m(\Omega_g)^* \beta_g \tag{1}$$

with ω the temporal frequency, β_g the sampling weights that can be calculated depending on the grid type, and the G discrete HRTF-angles $\Omega_g = \{(\phi_1, \theta_1), \ldots, (\phi_G, \theta_G)\}$ at azimuth ϕ, and elevation θ. The notation $(\cdot)^*$ denotes complex conjugation, and Y_n^m the spherical harmonics of order n and mode/degree m

$$Y_n^m(\theta, \phi) = \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{im\phi} \tag{2}$$

with the associated Legendre functions P_n^m [19, ch. 6.3], [20, ch. 1.3], and $i = \sqrt{-1}$ the imaginary unit. The inverse spatial Fourier transform can be used to recover H at arbitrary angles

$$\widehat{H}(\omega, \Omega) = \sum_{n=0}^{N} \sum_{m=-n}^{n} f_{nm}(\omega) Y_n^m(\Omega) \tag{3}$$

where N denotes the maximal order. If H is strictly order-limited, a sufficient choice of N results in $H = \widehat{H}$.

Depending on the spatial sampling grid Ω_g, the coefficients f_{nm} can be calculated up to a maximum order N

$$G \approx \eta (N+1)^2 \tag{4}$$

with η representing the efficiency of the sampling grid. The Lebedev grid [32], which will be used in the following, achieves $\eta = 1.3$ [4]. In case the order of H exceeds N, spatial aliasing occurs [20, ch. 3.7]. In this context, an appropriate preprocessing that reduces the spatial complexity of H will directly relax the requirement on G.

The following paragraph describes the SUpDEq method, which performs a spatial upsampling of a sparse HRTF set according to the block diagram in Fig. 1. The entire processing is identical for the left and right ear signals, and corresponding subscripts were omitted in the following for ease of display. In a first step, the sparse HRTF set H_{HRTF} measured at S sampling points $\Omega_s = \{(\phi_1, \theta_1), \ldots, (\phi_S, \theta_S)\}$ is equalized with an appropriate equalization dataset H_{EQ}

$$H_{\text{HRTF,EQ}}(\omega, \Omega_s) = \frac{H_{\text{HRTF}}(\omega, \Omega_s)}{H_{\text{EQ}}(\omega, \Omega_s)}. \tag{5}$$

The equalization dataset is intended to remove the directional dependency in H_{HRTF} to a certain degree with the goal to minimize the required order for the SH transform. Different equalization datasets can be applied – throughout this study a rigid sphere transfer function is used [19, p. 227]

$$H_{\text{STF}}(\omega, \Omega_g) = P 4\pi \sum_{n=0}^{N_{\text{high}}} \sum_{m=-n}^{n} i^n j_n(kr) Y_n^m(\Omega_e) Y_n^m(\Omega_g)^* \tag{6}$$

with j_n the spherical Bessel function of the first kind, and the ear position Ω_e at $\phi = \pm 90^\circ$ and $\theta = 0^\circ$. P denotes an arbitrary sound pressure. The radius r should match the physical dimensions of a human head. The STF can, thus, be regarded as a simplified HRTF set which features basic temporal and spectral components but does not carry information on the shape of the outer ears nor the fine structure of the head. The equalization with the STF indeed considerably reduces the spatial order of H_{HRTF}, and thus also the required maximal order N and number of sampling points S, as will be shown in the evaluation. Except for an angular shift in azimuth, the set is identical for the left and the right ear. As the equalization dataset is based on an

analytical description, it can be determined at a freely chosen maximal order, typically, a high order $N_{\text{high}} \geq 35$.

Equalization datasets based on other models, such as an ellipsoid for example, may also be appropriate, but using a rigid sphere model as the basis of the equalization dataset has several advantages. First, rigid sphere models have been extensively studied regarding their suitability for modeling a human head [33], [34]. Furthermore, for example Algazi *et al.* [35] presented a method to determine the optimal sphere radius based on the anthropometry of the actual head to be modeled. Applying this, customized STFs fitted to the respective (human) head can be obtained. Thus, rigid sphere models as well as various optimization methods are well established in research, whereas descriptions of other models are comparably rare and less convenient to implement.

In a second step, SH coefficients $f_{\text{EQ,nm}}$ for the equalized sparse HRTF set are obtained by applying the SH transform according to (1) on the equalized HRTFs given by (5) up to an appropriate low maximal order N_{low} that satisfies (4). In a third step, an upsampled HRTF set $\widehat{H}_{\text{HRTF,EQ}}$ is calculated on a dense sampling grid $\Omega_d = \{(\phi_1, \theta_1), \ldots, (\phi_D, \theta_D)\}$, with $D \gg S$ by using the inverse SH transform described by (3). In a fourth and final step, HRTFs are reconstructed by a subsequent de-equalization by means of spectral multiplication with a de-equalization dataset H_{DEQ}

$$\widehat{H}_{\text{HRTF,DEQ}}(\omega, \Omega_d) = \widehat{H}_{\text{HRTF,EQ}}(\omega, \Omega_d) \cdot H_{\text{DEQ}}(\omega, \Omega_d)\,. \quad (7)$$

This last step recovers energies at higher spatial orders that were transformed to lower orders in the first step. For de-equalization, the STF as given in (6) is used in the present study, whereas alternative choices will be addressed in the conclusion.

Again, $H_{\text{HRTF}} = \widehat{H}_{\text{HRTF,DEQ}}$ holds if N_{low} and N_{high} are chosen appropriately. Otherwise, deviations will be caused by signal energy which, after the equalization, still is apparent at high modal orders $N > N_{\text{low}}$. Due to spatial aliasing, this signal energy is irreversibly mirrored to lower orders $N \leq N_{\text{low}}$ [21], and we obtain $H_{\text{HRTF}} \approx \widehat{H}_{\text{HRTF,DEQ}}$. The following section analyzes the influence of these deviations, and which advantage the SUpDEq method provides in comparison to common order-limited SH interpolation without equalization.

III. Evaluation

We implemented the SUpDEq method in MATLAB utilizing routines from SOFiA toolbox [36] and AKtools [37] for spherical harmonics signal processing. We used the standardized AES69 file format [38] for in- and output of the HRTFs. Optionally, HRTFs can be passed in the frequency or the SH domain. Although the SUpDEq method might be most advantageous for reducing the measurement effort of individual HRTF sets, we used HRTFs of a Neumann KU100 dummy head that were measured on a Lebedev grid with 2702 sampling points and can be used for SH processing up to $N = 35$ for an initial evaluation [6]. Using dummy head datasets is common (see, e.g., [16], [21], [24]–[26], [28]) and necessary because the evaluation requires high order SH processing, which is hardly possible with any public database of individual HRTFs due to their irregular non full-spherical spatial sampling grids.

To generate various sparse HRTF sets required as input data for the evaluation, we simply spatially subsampled the reference set in the SH domain by means of the inverse SH transform. The (optimal) radius for the rigid sphere model was calculated according to Algazi *et al.* [35] based on the dimensions of the Neumann KU100 dummy head. With a head width of 15.5 cm, a head height of 25 cm, and a head length of 20 cm, the resulting radius was $r = 9.19$ cm.

A. Effect on Spherical Harmonics Representation

First, we analyzed how the equalization and the de-equalization affect the energy distribution across the different spatial modes. For this, we equalized the reference HRTF set on the corresponding dense Lebedev grid Ω_d (abbreviated Lebedev$_{2702}$ in the following) according to (5) (in contrast to the usual procedure where the equalization is performed on a sparse grid Ω_s) and transformed this equalized reference HRTF set to the SH domain with $N = 35$ according to (1). Next, we exemplarily applied the entire processing chain as described in Section II to a sparse HRTF set on a Lebedev grid Ω_s with only 38 sampling points ($N_{\text{low}} = 4$). Here, the de-equalization according to (7) was performed on the same dense Lebedev$_{2702}$ grid Ω_d. Likewise, the final de-equalized HRTF set was transformed to the SH domain with $N = 35$ to be able to compare the different outcomes.

Fig. 2 shows the corresponding magnitudes over the different spatial modes for the reference HRTF set $f_{\text{REF,nm}}$, the equalized reference set $f_{\text{EQ,nm}}$, and the de-equalized example set $f_{\text{DEQ,nm}}$. As expected, for frequencies above 8 kHz, $f_{\text{REF,nm}}$ shows considerable magnitudes even at high orders $N \geq 20$ [see Fig. 2(a)]. The equalized counterpart $f_{\text{EQ,nm}}$, however, contains almost no remarkable magnitudes at orders of $N \geq 5$ [see Fig. 2(b)]. Thus, the equalization successfully reduces the spatial complexity by eliminating most of the components at high orders. In theory, the better the rigid sphere model fits the human head (or in this case the dummy head), the more components at higher orders are reduced by the equalization. Ideally, only components at $N = 0$ would remain. In contrast to the equalized set, the de-equalized set $f_{\text{DEQ,nm}}$ again contains relevant components even at high orders $N \geq 20$ [see Fig. 2(c)]. Thus, the de-equalization recovers most of the basic magnitude structure that can also be found in the reference set $f_{\text{REF,nm}}$, even though these components were not inherent in the original sparse HRTF set with 38 sampling points and $N_{\text{low}} = 4$. However, some components are not appropriately reconstructed, and overall $f_{\text{REF,nm}}$ is more finely structured than $f_{\text{DEQ,nm}}$. These differences might result from the estimation error applying the rigid sphere model [11]. Moreover, the more sampling points the sparse set contains, the smaller are the differences between the reference and the de-equalized set. This relationship between spatial resolution of the sparse set and reconstruction accuracy will be further addressed in Section III-B.

Fig. 3 provides another perspective on how the equalization affects the energy distribution over N. The energy per order was calculated using Parseval's theorem [20, p. 18] and normalized by $1/(4\pi)$. As can be seen, the energy of the equalized set $f_{\text{EQ,nm}}$ is already substantially decreased at lower orders of about $N = 6$, whereas the reference set $f_{\text{REF,nm}}$ contains almost the same

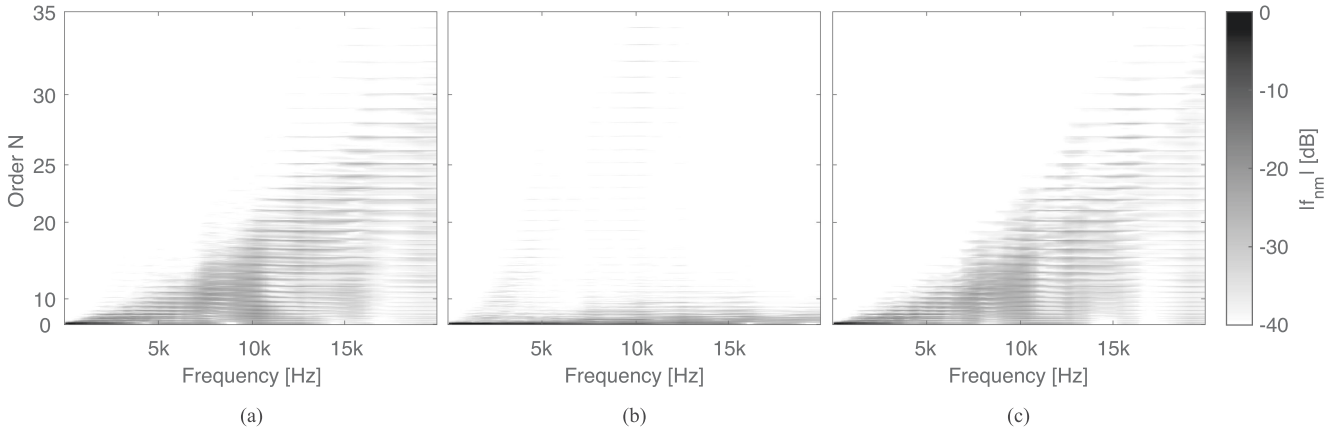


Fig. 2. Distribution of magnitudes over the different spatial modes for the left ear, orders up to $N = 35$, and frequencies up to $f = 20$ kHz. (a) Reference set $f_{\text{REF,nm}}$. (b) Equalized reference set $f_{\text{EQ,nm}}$. (c) De-equalized example set $f_{\text{DEQ,nm}}$, based on a sparse HRTF set with only 38 sampling points ($N_{\text{low}} = 4$).

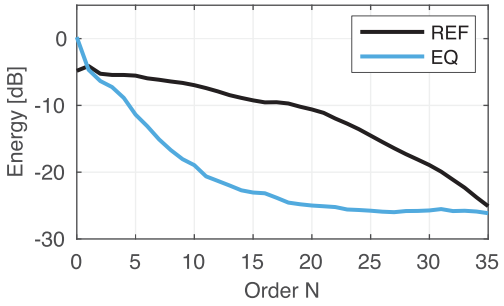


Fig. 3. Energy of the reference set $f_{\text{REF,nm}}$ (black) and the equalized reference set $f_{\text{EQ,nm}}$ (blue) as a function of order N.

amount of energy per order up to $N = 10$ before the energy starts to decrease slightly. For example, in $f_{\text{EQ,nm}}$, the energy is already decreased by more than 12 dB at $N = 6$ (0 dB at $N = 0$, -13 dB at $N = 6$). In contrast, the energy in $f_{\text{REF,nm}}$ is decreased by about 12 dB only at a much higher order $N = 28$ (-5 dB at $N = 0$, -17 dB at $N = 28$). Thus, as already outlined in the previous paragraph, the equalization successfully eliminates components at higher orders and thereby reduces spatial complexity.

B. Effect on Spatially Upsampled HRTFs

In a next step, we compared HRTFs obtained with the SUpDEq method (de-equalized HRTFs) to HRTFs obtained with the SH interpolation without any pre- or postprocessing before or after the SH transform. As discussed in the introduction, there are various pre- and postprocessing techniques which can be applied to the sparse HRTF set or to the SH-interpolated HRTF to improve the results (see, e.g., Brinkmann and Weinzierl [28] for a comparison). However, our intention was to compare the SUpDEq method to a well known and accepted procedure. For this, we generated SH coefficients from 15 sparse sampling grids – Lebedev grids with 6, 14, 26, 38, 50, 74, 86, 110, 146, 170, 194, 230, 266, 302, and 350 sampling points – which equals (limited) orders of $N = 1 - 15$. The coefficients will be referred to as $f_{\text{DEQ,nm}}$ and $f_{\text{OL,nm}}$ with OL standing for (strictly) order-limited. Thus, both order-limited and de-equalized (DEQ) sets were always based on the respective sparse grid. HRTFs were then obtained via the inverse SH transform at the positions required for the different evaluation methods (see below).

1) Impulse Responses and Transfer Functions: To get a first impression of the results, we compared the reference HRTFs (and the respective HRIRs) to HRTFs obtained with OL interpolation or with the SUpDEq method. As an example, we extracted left ear HRTFs for the frontal ($\phi = 0°, \theta = 0°$) and the more critical contralateral direction ($\phi = 270°, \theta = 0°$) by means of SH interpolation based on a sparse HRTF set sampled on a Lebedev grid Ω_s ($N_{\text{low}} = 4$, 38 sampling points). Fig. 4 illustrates the respective magnitude and impulse responses.

For the frontal direction, the magnitude response of the de-equalized HRTF is in good agreement with the reference, apart from a slight ripple above approximately 2 kHz. In contrast, the order-limited HRTF clearly suffers from the typical high-frequency deviations for the frontal direction due to spatial aliasing (see, e.g., [21], [24], [25]). For the contralateral direction, however, the magnitude responses of both HRTFs show distinct distortions above 1 kHz.

An inspection of the HRIRs reveals another interesting point: Whereas the temporal structure of the de-equalized HRIR follows that of the reference – especially for the frontal direction – the order-limited HRIR shows distinct pre-ringing artifacts. The preringing occurs for both directions, but for contralateral sound incidence it spreads over a time period that even exceeds the visible length of the reference HRIR. This preringing most likely deteriorates the ITDs in this case, which is less likely the case for the de-equalized impulse responses. Whereas this is admittedly an extreme example due to the low order of $N_{\text{low}} = 4$, it already illustrates the potential of the SUpDEq method. Moreover, informal inspections of results from higher orders showed that the observed ringing artifacts in the order-limited HRIRs decrease with order but remain visible even at $N_{\text{low}} = 15$.

In general, contralateral directions are much more critical than frontal or ipsilateral directions, mostly because diffraction around the head influences sound incidence at the contralateral ear and is not adequately matched by the applied (de-)equalization dataset. Obviously, the differences for contralateral directions also decrease with increasing order N_{low} of the sparse set. The effects observed for contralateral directions are discussed in more detail in Section IV.

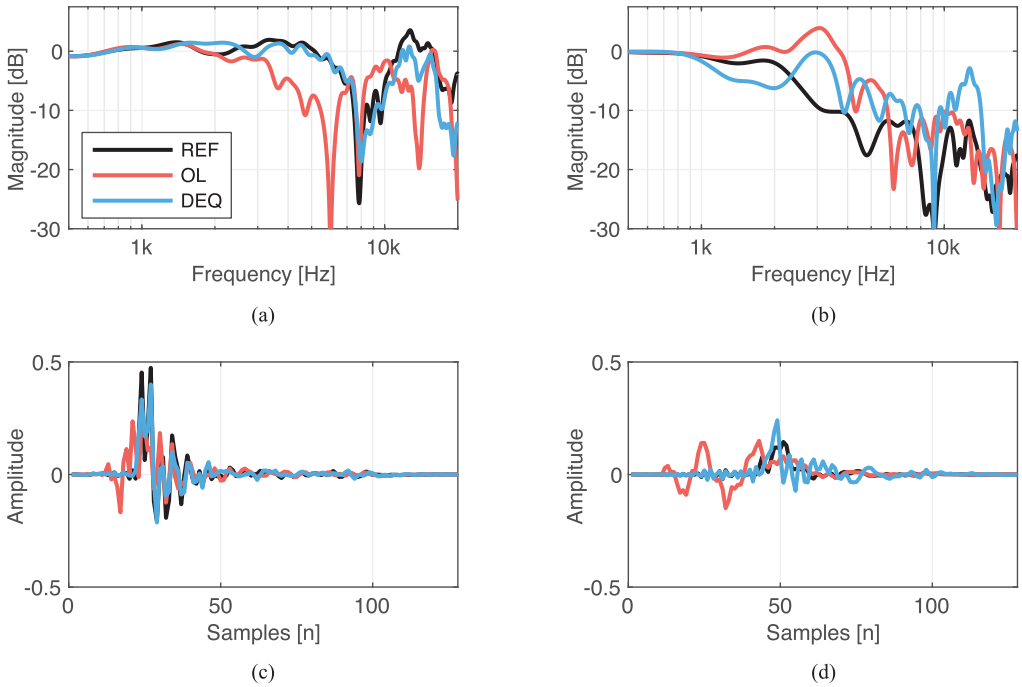


Fig. 4. Left ear magnitude (top) and impulse (bottom) responses, extracted from the reference set $f_{\text{REF,nm}}$ (black), the order-limited set $f_{\text{OL,nm}}$ (red), and the de-equalized set $f_{\text{DEQ,nm}}$ (blue) by SH interpolation [$f_{\text{OL,nm}}$ and $f_{\text{DEQ,nm}}$ based on a sparse HRTF set with 38 sampling points ($N_{\text{low}} = 4$)]. (a), (c) Front direction ($\phi = 0°, \theta = 0°$). (b), (d) Contralateral direction ($\phi = 270°, \theta = 0°$).

2) Spectral Differences: Next, we analyzed the spectral deviations to the reference set $f_{\text{REF,nm}}$ as a function of N on various test sampling grids with T sampling points $\Omega_t = \{(\phi_1, \theta_1), \ldots, (\phi_T, \theta_T)\}$. The frequency-dependent spectral differences per sampling point were calculated in dB as

$$\Delta g(\omega, \Omega_t) = 20lg \frac{\mid H_{\text{HRTF,REF}}(\omega, \Omega_t) \mid}{\mid H_{\text{HRTF,TEST}}(\omega, \Omega_t) \mid} \tag{8}$$

where $H_{\text{HRTF,REF}}$ is the left ear HRTF extracted from $f_{\text{REF,nm}}$ and $H_{\text{HRTF,TEST}}$ is the left ear HRTF extracted from $f_{\text{OL,nm}}$ or $f_{\text{DEQ,nm}}$ at the sampling point Ω_t. Then, the absolute value of $\Delta g(\omega, \Omega_t)$ was averaged across the temporal frequency ω to obtain one value $\Delta G_{sp}(\Omega_t)$ (in dB) per sampling point

$$\Delta G_{\text{sp}}(\Omega_t) = \frac{1}{N_\omega} \sum_{\omega} \mid \Delta g(\omega, \Omega_t) \mid \tag{9}$$

across all sampling points Ω_t to obtain the frequency-dependent measure $\Delta G_f(\omega)$ (in dB)

$$\Delta G_f(\omega) = \frac{1}{N_{\Omega_t}} \sum_{\Omega_t} \mid \Delta g(\omega, \Omega_t) \mid \tag{10}$$

and across ω and Ω_t, resulting in a single value ΔG (in dB) describing the spectral difference

$$\Delta G = \frac{1}{N_{\Omega_t}} \frac{1}{N_\omega} \sum_{\Omega_t} \sum_{\omega} \mid \Delta g(\omega, \Omega_t) \mid . \tag{11}$$

Fig. 5(a) shows the spectral differences ΔG over the full audio bandwidth across N for OL interpolation and the SUpDEq method. The test sampling grid Ω_t was the reference Lebedev$_{2702}$ grid. As can be seen, the SUpDEq method outperforms the OL interpolation in all conditions. The spectral differences for the SUpDEq method are always about 2 dB lower than for OL interpolation. Thus, already at $N = 4$, the SUpDEq method leads to spectral differences of about 3 dB, and at $N \geq 9$, the differences are constantly around only 2 dB. In contrast, the differences for OL interpolation are never below 4 dB.

Fig. 5(b) illustrates the frequency-dependent spectral differences $\Delta G_f(\omega)$ at $N = 4, 7, 10, 13$ for the SUpDEq method and for OL interpolation. Just like above, the test sampling grid Ω_t was the Lebedev$_{2702}$ grid. When looking at the plot, two things can be observed. First, as expected based on ΔG in Fig. 5(a), the spectral differences are distinctly smaller for the SUpDEq method than for OL interpolation. Second, in comparison to the SUpDEq method, the spectral differences for OL interpolation exceed 2 dB already at lower frequencies (between 2 and 6 kHz, depending on N) and, above this specific alias frequency, suddenly rise from 2 dB up to about 4 to 6.5 dB within about one octave. Then, at frequencies above 10 kHz, the differences are constantly high at about 5 to 7 dB. For the SUpDEq method, however, the spectral differences show a much more gentle rise. Here, the differences exceed 2 dB at frequencies between 3 and 9 kHz (depending on N) and mostly stay between only 2 and 4 dB, even at higher frequencies above 10 kHz. Thus, with the SUpDEq method, already at $N = 10$ the spectral differences hardly exceed 2 dB over the entire audio bandwidth.

Fig. 5(c) and (d) concludes the spectral analysis. These plots show the spectral differences $\Delta G_{\text{sp}}(\Omega_t)$ per sampling point for OL interpolation [see Fig. 5(c)] and for the SUpDEq method [see Fig. 5(d)] at $N = 4$ and $f \leq 10$ kHz. In this case, the test sampling grid Ω_t was full spherical with ϕ from 0° to 359°

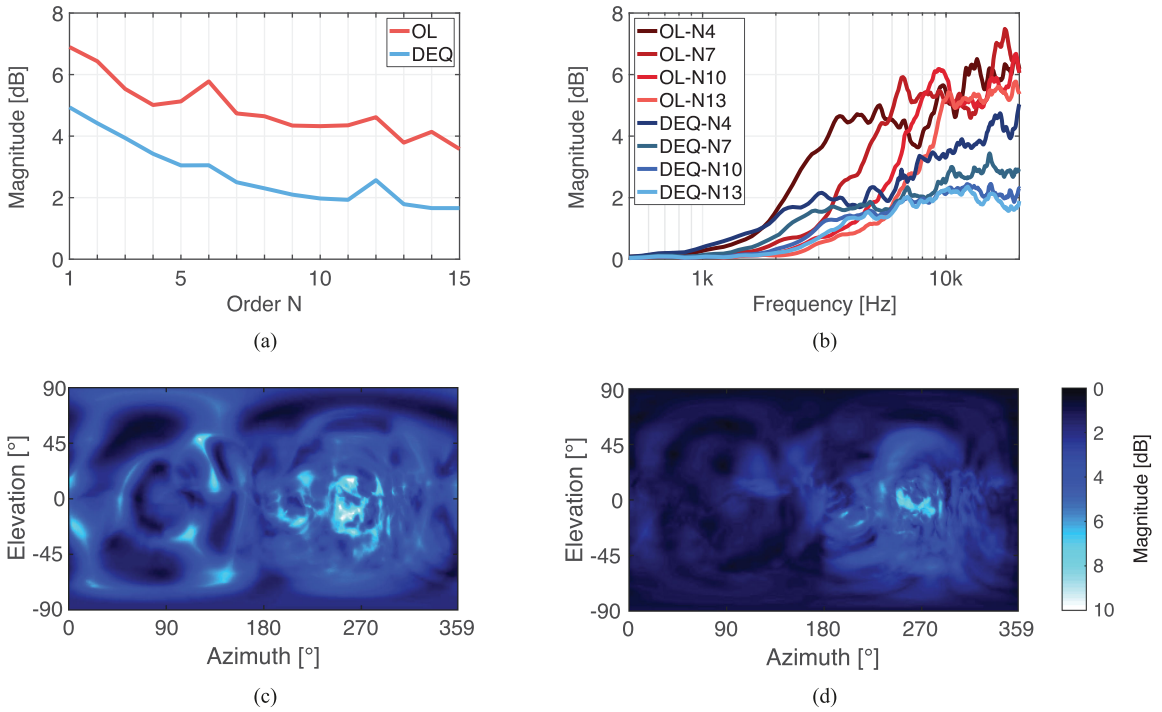


Fig. 5. Spectral differences in dB (left ear) between reference HRTF sets (extracted from $f_{\mathrm{REF,nm}}$) and order limited or de-equalized HRTF sets (extracted from $f_{\mathrm{OL,nm}}$ or $f_{\mathrm{DEQ,nm}}$, both based on the respective sparse HRTF set). (a) Spectral differences ΔG over the full audio bandwidth across N for OL interpolation (red), the SUpDEq method (blue), and Lebedev$_{2702}$ sampling grid. (b) Frequency-dependent spectral differences $\Delta G_f(\omega)$ for $N = 4, 7, 10, 13$ (color saturation) and Lebedev$_{2702}$ sampling grid. Spectral differences $\Delta G_{sp}(\Omega_t)$ per sampling point for OL interpolation (c) and for the SUpDEq method (d) at $N = 4$ and $f \leq 10$ kHz (see text for applied sampling grids).

and θ from $-90°$ to $+90°$, each in steps of $1°$, leading to a total of 65160 sampling points. The plots reveal that OL interpolation results in distinct spectral differences spread over the entire angular range. In contrast, the SUpDEq method leads to spectral differences mainly located at contralateral directions. As already mentioned, the artifacts observed for contralateral directions are discussed in more detail in Section IV. At the front direction, where OL interpolation typically performs badly, the SUpDEq method shows extremely good results. The same can be observed for various ipsilateral directions. Despite the spatial distribution, the spectral differences are generally higher for OL interpolation, with a maximum of about $\Delta G_{\mathrm{sp}}(\Omega_t) = 11$ dB at $\phi = 264°$ and $\theta = 13°$. The SUpDEq method results in a maximum spectral difference $\Delta G_{\mathrm{sp}}(\Omega_t)$ of about 8 dB at $\phi = 252°$ and $\theta = 0°$.

3) Binaural Cues: In a further analysis, we compared the ILDs and ITDs – which are the dominant cues for left-right localization – of the reference HRTF set with those of order-limited and de-equalized sets. For this purpose, we extracted HRTFs in the horizontal plane ($\theta = 0°$) with an angular spacing of $\phi = 1°$ from the reference set $f_{\mathrm{REF,nm}}$ and, depending on N, from the respective order-limited and de-equalized set $f_{\mathrm{OL,nm}}$ and $f_{\mathrm{DEQ,nm}}$. The broadband ILDs were then calculated as the ratio between the energy of the left and right ear HRIR. The ITDs were calculated by means of a threshold-based onset detection on the ten times up-sampled and low-pass filtered HRIRs (10th order Butterworth low pass at 3 kHz). This extraction method showed the best agreement with the perceived source position [39], and reflects that the auditory system mainly exploits the ITD for frequencies below approximately 1.5 kHz [1]. As a side effect, the low-pass filter also suppressed the strong preringing in the order-limited HRIRs that can be seen in Fig. 4(c) and (d), which might indicate that this effect is of less perceptual relevance.

Fig. 6 shows the calculated ILDs and ITDs of the reference HRTF set as well as of the order-limited and de-equalized sets for $N = 4, 7, 10, 13$. As can be seen in Fig. 6(a), the ILDs of the de-equalized HRTFs are in good agreement with the ILDs of the reference for all tested spatial orders. Surprisingly, the deviations from the reference do not decrease monotonically with increasing order. Thus, the ILDs differ most from the reference at $N = 4$, but also at $N = 13$, especially due to the deviations at lateral directions. At $N = 7$ and $N = 10$, though, the ILDs of the de-equalized HRTFs correspond approximately to the reference. In contrast, the ILDs of the order-limited HRTFs presented in Fig. 6(c) deviate strongly from the reference over the entire angular range and generally vary clearly with respect to N. As expected, the differences become smaller with increasing spatial order, but even at $N = 13$, distinct differences remain.

The ITDs of the de-equalized HRTFs shown in Fig. 6(b) are almost identical to the ITDs of the reference across all tested spatial orders. Here, only marginal deviations can be observed at lateral and rearward directions. The ITDs of the order-limited HRTFs, though, differ clearly from the reference at $N = 4$, but also are almost equal to the reference at $N = 7, 10, 13$, with only slight deviations at the same lateral and rearward directions [see Fig. 6(d)]. However, it must be emphasized that informative ITDs of the order-limited HRTFs presented here could only be determined because the high-frequency preringing was

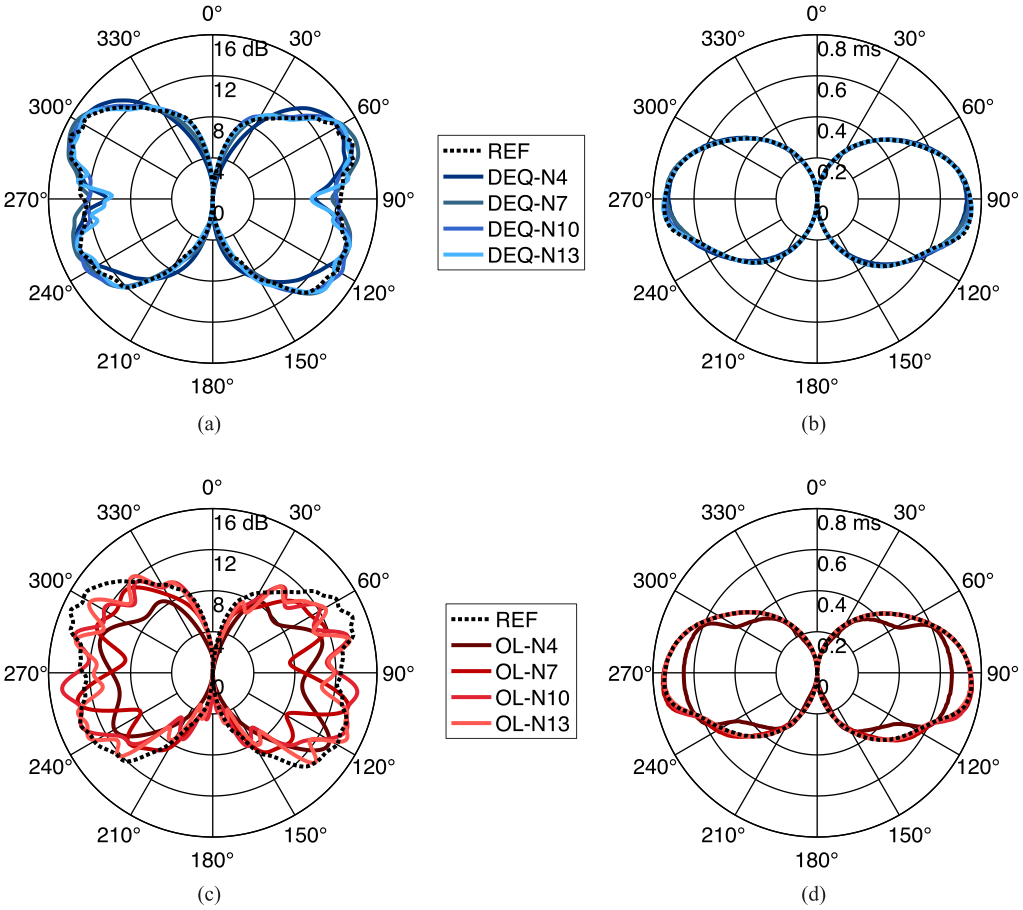


Fig. 6. ILDs (a), (c) and ITDs (b), (d) in the horizontal plane for the reference (black) HRTF set (extracted from $f_{\mathrm{REF,nm}}$) and for order-limited (red) or de-equalized (blue) HRTF sets (extracted from $f_{\mathrm{OL,nm}}$ or $f_{\mathrm{DEQ,nm}}$, both based on the respective sparse HRTF set) for $N = 4, 7, 10, 13$ (color saturation). The angle represents the azimuth ϕ of the sound source. The radius describes the magnitude of the level differences (in dB) or time differences (in ms).

previously filtered out. The extent to which preringing actually influences ITDs and, thus, the localization has to be further investigated.

4) Localization Performance: To conclude the evaluation, we compared the localization performance of order-limited HRTFs and de-equalized HRTFs with two different auditory models. To assess the localization performance in the median sagittal plane, we used the model from Baumgartner *et al.* [40]. This model compares the spectral structure of a reference HRTF set to a set of test HRTFs to calculate a probabilistic estimate of the perceived sound source location. Based on this, it calculates the polar RMS error (PE, in degrees) describing the expected angular error between the actual and perceived source positions as well as the quadrant error rate (QE, in percent) specifying the rate of front back or up-down confusions. To estimate the performance in the horizontal plane, we applied the model from May *et al.* [41]. This probabilistic model is based on a trained Gaussian mixture model that weights the frequency-dependent binaural cues (ILDs, ITDs) to estimate the azimuthal position of a sound source. By comparing the intended and the estimated source position, a lateral error (LE, in degrees) can be calculated. Both models are part of the auditory modeling toolbox [42], which we used for the analysis.

The procedure was as follows: we first determined the performance of the reference set $f_{\mathrm{REF,nm}}$, the order-limited set $f_{\mathrm{OL,nm}}$, and the de-equalized set $f_{\mathrm{DEQ,nm}}$ as a function of N. To estimate median plane localization performance, we used a test sampling grid Ω_t with $\phi = \{0°, 180°\}$ and $-30° \leq \theta \leq 90°$ in steps of $1°$, and assumed a median listener sensitivity of $S = 0.76$ (in accordance with Baumgartner *et al.* [40]). To estimate the horizontal plane localization performance, we applied a test sampling grid with $\phi = \pm 90°$ in steps of $5°$. We then calculated the absolute polar error difference (in degree)

$$\Delta\mathrm{PE} = \mid \mathrm{PE}_{\mathrm{REF}} - \mathrm{PE}_{\mathrm{TEST}} \mid \tag{12}$$

the absolute quadrant error difference (in percent)

$$\Delta\mathrm{QE} = \mid \mathrm{QE}_{\mathrm{REF}} - \mathrm{QE}_{\mathrm{TEST}} \mid \tag{13}$$

as well as the absolute lateral error difference (in degree)

$$\Delta\mathrm{LE} = \frac{1}{T}\sum_{t=1}^{T} \mid \mathrm{LE}_{\mathrm{REF}}(\Omega_t) - \mathrm{LE}_{\mathrm{TEST}}(\Omega_t) \mid \tag{14}$$

for each order N with the subscripts REF and TEST as defined above. In some cases where the error of the test condition was less than the error of the reference, we set the absolute error to 0. Thus, all three measures (ΔPE, ΔQE, and ΔLE) describe how the already existing localization errors of the reference HRTF set change if order-limited or de-equalized HRTF sets are applied alternatively.

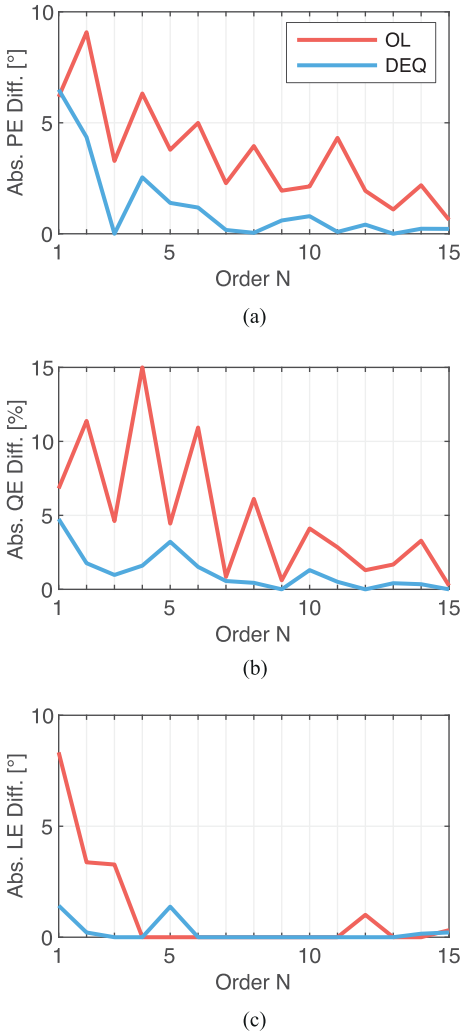


Fig. 7. Absolute polar error difference ΔPE (a), quadrant error difference ΔQE (b), and lateral error difference ΔLE (c) over SH order N for OL interpolation (red) and the SUpDEq method (blue).

In the median sagittal plane [see Fig. 7(a) and (b)], the OL interpolation leads to considerably higher errors over the entire range of N. Obviously, the high-frequency deviations found in order-limited HRTFs (the result of the OL interpolation) badly affects spectral cues that are responsible for median sagittal plane localization. As expected, ΔPE decreases with increasing order N, approximating ΔPE $\leq 2°$ at $N \geq 12$. The extent of the quadrant error increase varies greatly and lies between 5 and 15 % at $N \leq 6$, reaching values between 5 and 2 % at $N \geq 7$. The SUpDEq method, however, amplifies the polar error only slightly at lower orders $N \leq 2$. At higher orders, ΔPE is always between 2 and 0°, indicating that the spectral cues are marginally impaired at $N \leq 6$ and mostly unimpaired at $N \geq 7$. In accordance with this, ΔQE is always below 5 % and reaches values around 0 % at $N \geq 7$.

In the horizontal plane [see Fig. 7(c)], the OL interpolation performs a little better, although lateral errors are distinctly amplified at orders $N \leq 3$. The reason for the increased errors are most certainly the strong preringing artifacts causing wrong ITDs, as already shown in Section III-B1. At orders $N \geq 4$, the additional error levels off at about 1°, even though the HRIRs still have notable preringing (see Fig. 4) at $N = 4$. In contrast, the SUpDEq method leads to hardly any increase in lateral error over the entire tested range of N (ΔLE averaged over N of about 0.2°).

IV. Discussion

In the evaluation, we analyzed the proposed directional (de-)equalization in detail by comparing HRTFs obtained with the SUpDEq method to HRTFs obtained with common OL interpolation regarding their representation in the SH domain, their spectral and temporal structure, their binaural cues, and their localization performance in the horizontal and median sagittal planes. We showed that the proposed directional equalization successfully decreases the spatial complexity of an HRTF set by reducing the energy in higher SH orders. As a result, an equalized set can be sufficiently described by lower orders. The (de-)equalization on a dense grid, on the other hand, successfully recovers relevant components at higher orders which were not inherent in the sparse HRTF set. The accuracy of the (de-)equalization depends on the number of sampling points of the sparse HRTF set and on the extent to which the applied equalization dataset fits the human head.

Regarding the spectral structure, the deviations from the reference HRTF set are significantly smaller for the SUpDEq method than for OL interpolation. Averaged over frequency and all sampling points, depending on N of the sparse HRTF set, the spectral differences averaged over frequency to the reference set are only about 5 dB at $N = 1$ and already below 2 dB at $N \geq 9$ for the SUpDEq method. In contrast, the differences for OL interpolation are still around 4 dB at $N = 15$. Considering the spectral differences over frequency, the SUpDEq method also performs much better. In comparison, the differences show a much more gentle rise over frequency with maximum values of about 2 to 4 dB ($N = 13, 10, 7, 4$) at frequencies above 10 kHz. The spectral differences for OL interpolation, though, increase sharply after the alias frequency and stay constantly high at about 5 to 7 dB ($N = 13, 10, 7, 4$) at frequencies above 10 kHz. Moreover, spectral differences induced by the SUpDEq method are mainly located at contralateral directions, whereas the (stronger) spectral differences caused by OL interpolation are spread over the entire angular range, with distinct clusters at frontal and contralateral directions.

The results of the spectral and temporal analysis are also reflected in the modeled localization performance. Here, the SUpDEq method performed better in both planes (the median sagittal plane and the horizontal plane) because spectral and binaural cues are not impaired as much as with OL interpolation. Thus, the SUpDEq method caused hardly any differences in polar errors at $N \geq 5$ and in lateral errors over the entire tested range of N as compared to the reference. In contrast, especially in the median sagittal plane, the OL interpolation caused a distinct increase in error, even up to higher orders.

Throughout the entire evaluation, we recognized that independent of the upsampling method, the differences to the reference are maximal for contralateral directions, both for the temporal structure as well as for the spectrum. This can be explained by the propagation of sound around the head or a rigid sphere to

the contralateral far side. Here, a bright spot can be observed because all waves reach the far side in phase and, thus, interfere constructively [34]. However, the interference pattern changes rapidly for adjacent directions, especially towards higher temporal frequencies. This corresponds to a high spatial order N required to cover these changes. Buchanan *et al.* [43] for example found similar difficulties to reconstruct the STF at contralateral regions with a filter model, especially because of strong changes of the STF at only small spatial shifts, which is in line with our observations. Thus, when describing the sound incidence by sparse grids, due to the limited orders, differences to the reference remain and tend to be maximal for the contralateral side. A directional equalization of the HRTF set as proposed in the present paper reduces the energy in higher spatial orders, but only to the extent to what the equalization function (STF) matches the properties of the human head. As the rigid sphere model corresponds only approximately to a human head, different interference patterns occur for a sphere and a human head. The differences remaining after a directional equalization are maximal for sound waves reaching the area of the contralateral side, and thus the spectral differences of the de-equalized set become maximal for these directions as well. According to this, to appropriately cover the contralateral directions, a higher spatial order N is required than for the other directions.

The observed temporal and spectral errors at contralateral regions have only a minor influence on the modeled localization performance (see. Fig. 7). Horizontal plane localization is dominated by the low-frequency ITD [44], which is reconstructed almost perfectly by the SUpDEq method in all cases, whereas slightly worse results of the OL approach can be explained by the less accurate ITD reconstruction at low SH orders [see Fig. 6(b) and (d)]. The median sagittal plane localization on the other hand relies on high-frequency spectral cues, which are well reconstructed by the SUpDEq method already at relatively low SH orders, at least at noncontralateral directions (see Figs. 4 and 5). Thus, the modeled localization error increases only slightly at low orders for de-equalized HRTFs, whereas larger errors of the order-limited HRTFs stem from the severe high-frequency coloration that can be observed almost regardless of the source position and up to relatively high SH orders. This coloration also affects the ILD of the order-limited HRTFs that–in contrast to the de-equalized data–shows relatively large errors for most source azimuths and up to high SH orders. Moreover, the importance of the contralateral ear for median plane localization decreases with increasing source lateralization [45], which the applied model considers by means of a binaural weighting factor [40]. Taking this into account, it can be hypothesized that median plane localization performance will be relatively constant across source azimuths. This, however, requires further validation.

Finally, it needs to be discussed to what extent the results found in this study can be generalized. Thus, we repeated the complete evaluation for another reference, namely the HRTF set of the FABIAN head and torso simulator [7]. Unlike the Neumann KU100, FABIAN has shoulders and a torso, and thus the HRTFs show even more characteristics typical for human HRTFs. For the sake of readability, we did not include the results of the additional evaluation in the paper, but attached them as supplementary materials. When applying SUpDEq to the FABIAN head, we observed comparable deviations from the reference and a similar localization performance as with the Neumann KU100 dummy head. Thus, the directional equalization of the FABIAN HRTFs results in a comparable reduction of the energy in higher orders. Moreover, the exemplarily considered HRTFs as well as the measurements of spectral differences showed a similar structure. This also applies to the analysis of the binaural cues (ILDs and ITDs) and to the modeled localization performance. For more details, please refer to the description and figures of the supplementary materials. Overall, we can conclude that the analysis with two exemplary dummy heads gives strong evidence that the results found in this study can be generalized to a variety of dummy heads, and probably also to individual HRTFs.

V. Conclusion

In current research, HRTF processing in the SH domain is widely discussed. In this study, we presented a SH-based approach for spatial upsampling (interpolation) of sparse HRTF sets and demonstrated that it can significantly reduce the measurement effort for obtaining high fidelity HRTF sets acquired in the laboratory or at home. The basic idea of the SUpDEq method is to remove direction-dependent temporal and spectral components of the HRTFs, such as frequency-dependent ITDs and ILDs as well as elevation-dependent spectral features, before the SH transform. This is achieved by a directional equalization (spectral division) with an appropriate equalization dataset – for example a rigid sphere transfer function as used in this study. In general, such a directional equalization can be used to realize other interpolation techniques [28] as well, for example by choosing an equalization dataset, which only includes the phase or the amplitude. As the equalization significantly reduces the spatial order of the sparse set, the equalized HRTF set can then be transformed to the SH domain at a low order N_{low} with only minor spatial-aliasing artifacts. After spatial upsampling to an arbitrary (full spherical) dense sampling grid using interpolation by means of the inverse SH transform, the directional de-equalization (spectral multiplication) is applied to the HRTFs to restore the previously discarded spatial components, resulting in a dense set of conventional HRTFs.

The evaluation revealed that already with 38 HRTFs ($N_{\text{low}} = 4$), a decent full-spherical dense HRTF set (for example with 2702 sampling points, $N_{\text{max}} = 44$) can be generated. Dependent on direction, this set shows only small spectral and temporal deviations from a reference, and provides only slightly higher errors in localization performance than a reference in an auditory model based evaluation. Thus, the approach presented here can be regarded as one step closing the gap between a practical, fast, and simple measurement procedure and sufficient accuracy of the upsampled HRTF set.

Thus, as future work, we plan to investigate a larger number of individual datasets (see, e.g., [46]) to analyze if the results are similar for these HRTFs. Furthermore, as an important issue, we plan to evaluate the SUpDEq method with listening experiments to put the more technical results presented

in this paper in relation to auditory perception. For this, several test designs are imaginable. On the one hand, adaptive forced-choice procedures like QUEST [47] could be applied to estimate (direction-dependent) discrimination thresholds between reference HRTFs and de-equalized HRTFs, depending on N of the sparse input HRTF set. On the other hand, localization experiments could be performed to examine localization accuracy with respect to N. Furthermore, test paradigms like MUSHRA [48] or SAQI [49] could be used to assess quality-based parameters of de-equalized HRTFs in comparison to reference HRTFs. The SUpDEq method clearly aims at providing high quality individual HRTFs with a reduced measurement effort. However, depending on the application and the available SH order, dense nonindividual sets might still be preferable, which is another topic for future research. Finally, the presented method needs to be compared to other pre- and postprocessing methods [28] regarding technical and perceptual attributes.

Furthermore, since we have limited ourselves in this study to Lebedev grids of different orders, it might be worthwhile investigating other grid types (e.g., Gaussian grids, Fliege grids) as well. However, even though these grids might perform slightly different, we expect that the main results of the present study will be confirmed. The performance of the SUpDEq method could further be enhanced if a more exact description of the head geometry would be applied as the (de-)equalization function. Instead of a rigid sphere model, the head geometry could for example be described by an appropriate ellipsoid, which approximates a human head much better. The mathematical fundamentals have already been discussed [50], [51]. Alternatively, it might be possible to use a smoothed HRTF set [52] or a simplified set based on measurements as (de-)equalization dataset.

Using different equalization and de-equalization functions opens up new fields of application. For example, some adaptations of the SUpDEq method would provide the possibility to shift the sound source position form the far-field to the near-field and vice versa. The implementation is quite simple though. For the equalization, a (rigid sphere) head model describing the sound incidence of a far-field sound source in form of a plane wave is used, as done throughout this study. To obtain near-field HRTFs, a head model describing the sound incidence of a nearby point source (at a specific distance close to the head) needs to be applied for the de-equalization. The relevant equations describing the sound field of a point source reaching the human head can for example be found in Rafaely [20, ch. 2]. By this, and by considering the acoustic parallax effect of nearby sound sources, distance variations functions can be designed to generate near-field HRTF sets based on (sparse) far-field HRTF sets.

Individualization of HRTFs is another widely discussed topic in research (see, e.g., [46]). By using a different radius or ear position for the spherical head used during equalization and de-equalization, individual anthropometric features can be considered. This not only holds the possibility for individualizing the ITD, which was already given by previously suggested approaches (see Section I), but could also account for anthropometry related ILD changes and even asymmetries.

Finally, the approach could be transferred to sound sources. The principle of reciprocity [53] implies that the radiation from a distinct point on the sphere can be regarded in the same way as a sound wave reaching the sphere. Reciprocity has already been used to address comparable problems [54]–[57]. Thus, the methods proposed and described in this paper can as well be used to interpolate and analyze sound source directivities of speakers, instruments, or loudspeakers.

In this study, we have shown that an appropriate directional (de-)equalization of measured HRTF sets can be of great benefit for different fields of spatial audio and virtual acoustics. This paper has identified and assessed some of them and can be regarded as a starting point for further research. A MATLAB-based implementation of the SUpDEq method is available on https://github.com/AudioGroupCologne/SUpDEq.

REFERENCES

[1] J. Blauert, *Spatial Hearing - The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1996.

[2] M. Vorländer, *Auralization*. Berlin, Germany: Springer-Verlag, 2008.

[3] A. Roginska and P. Geluso, *Immersive Sound - The Art and Science of Binaural and Multi-Channel Audio*. New York, NY, USA: Routledge, 2018.

[4] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.

[5] W. G. Gardner and D. M. Keith, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Amer.*, vol. 97, no. 6, pp. 3907–3908, 1995.

[6] B. Bernschütz, "A spherical far field HRIR / HRTF compilation of the Neumann KU 100," in *Proc. 39th DAGA*, 2013, pp. 592–595.

[7] F. Brinkmann *et al.*, "A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations," *J. Audio Eng. Soc.*, vol. 65, no. 10, pp. 841–848, 2017.

[8] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?," *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, 1996.

[9] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. I: Stimulus synthesis," *J. Acoust. Soc. Amer.*, vol. 85, no. 2, pp. 868–878, 1989.

[10] V. R. Algazi, R. O. Duda, and D. M. Thompson, "The CIPIC HRTF database," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2001, pp. 99–102.

[11] R. Bomhardt and J. Fels, "Mismatch between interaural level differences derived from human heads and spherical models," in *Proc. 140th Audio Eng. Soc. Conv.*, Paris, France, 2016, pp. 1–10.

[12] A. Lindau and S. Weinzierl, "On the spatial resolution of virtual acoustic environments for head movements in horizontal, vertical and lateral direction," in *Proc. EAA Symp. Auralization*, 2009, pp. 1–6.

[13] E. M. Wenzel and S. H. Foster, "Perceptual consequences of interpolating head-related transfer functions during spatial synthesis," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 1993, pp. 102–105.

[14] J. Chen, B. D. Van Veen, and K. E. Hecox, "A spatial feature extraction and regularization model for the head-related transfer function," *J. Acoust. Soc. Amer.*, vol. 97, no. 1, pp. 439–452, 1995.

[15] E. H. A. Langendijk and A. W. Bronkhorst, "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 528–537, 2000.

[16] K. Hartung, J. Braasch, and S. J. Sterbing, "Comparison of different methods for the interpolation of head-related transfer functions," in *Proc. 16th Int. Audio Eng. Soc. Conf. Spatial Sound Reproduction*, 1999, pp. 319–329.

[17] T. Djelani, C. Pörschmann, J. Sahrhage, and J. Blauert, "An interactive virtual-environment generator for psychoacoustic research II: Collection of head-related impulse responses and evaluation of auditory localization," *Acta Acustica United Acustica*, vol. 86, no. 6, pp. 1046–1053, 2000.

[18] P. Minnaar, J. Plogsties, and F. Christensen, "Directional resolution of head-related transfer functions required in binaural synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 10, pp. 919–929, 2005.

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

[20] B. Rafaely, *Fundamentals of Spherical Array Processing*. Berlin Germany: Springer-Verlag, 2015.

[21] B. Bernschütz, A. V. Giner, C. Pörschmann, and J. M. Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica United Acustica*, vol. 100, no. 5, pp. 972–983, 2014.
[22] C. Pike, "Subjective assessment of HRTF interpolation with spherical harmonics," in *Proc. Int. Conf. Spatial Audio*, 2017. [Online]. Available: http://www.slideshare.net/ChrisPike21/subjective-assessment-of-hrtf-interpolation-with-spherical-harmonics-chris-pike-and-tony-tew, Accessed on: Apr. 2019.
[23] R. V. L. Hartley and T. C. Fry, "The binaural location of pure tones," *Phys. Rev.*, vol. 18, no. 6, pp. 431–442, 1921.
[24] M. Zaunschirm, C. Schoerkhuber, and R. Hoeldrich, "Binaural rendering of Ambisonic signals by HRIR time alignment and a diffuseness constraint," *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. 3616–3627, 2018.
[25] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, and S. Weinzierl, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *J. Acoust. Soc. Amer.*, vol. 141, no. 6, pp. 4087–4096, 2017.
[26] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Analyzing head-related transfer function measurements using surface spherical harmonics," *J. Acoust. Soc. Amer.*, vol. 104, no. 4, pp. 2400–2411, 1998.
[27] G. D. Romigh, D. S. Brungart, R. M. Stern, and B. D. Simpson, "Efficient real spherical harmonic representation of head-related transfer functions," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 921–930, Aug. 2015.
[28] F. Brinkmann and S. Weinzierl, "Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition," in *Proc. Audio Eng. Soc. Conf. Audio Virtual Augmented Reality*, 2018, pp. 1–10.
[29] H. Ziegelwanger and P. Majdak, "Modeling the direction-continuous time-of-arrival in head-related transfer functions," *J. Acoust. Soc. Amer.*, vol. 135, no. 3, pp. 1278–1293, 2014.
[30] R. Bomhardt, I. C. P. Mejia, A. Zell, and J. Fels, "Required measurement accuracy of head dimensions for modeling the interaural time difference," *J. Audio Eng. Soc.*, vol. 66, no. 3, pp. 114–126, 2018.
[31] H. Bahu and R. David, "Optimization and prediction of the spherical and ellipsoidal ITD model parameters using offset ears," in *Proc. Int. Audio Eng. Soc. Conf. Spatial Reproduction-Aesthetics Sci.*, 2018, pp. 1–11.
[32] V. I. Lebedev, "Spherical quadrature formulas exact to orders 2529," *Siberian Math. J.*, vol. 18, no. 1, pp. 132–142, 1977.
[33] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 6, no. 5, pp. 476–488, Sep. 1998.
[34] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Amer.*, vol. 104, no. 5, pp. 3048–3058, 1998.
[35] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a spherical-head model from anthropometry," *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 472–479, 2001.
[36] B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl, "SOFiA sound field analysis toolbox," in *Proc. Int. Conf. Spatial Audio*, 2011, pp. 8–16.
[37] F. Brinkmann and S. Weinzierl, "AKtools - an open software toolbox for signal acquisition, processing, and inspection in acoustics," in *Proc. 142nd Audio Eng. Soc. Conv.*, Berlin, Germany, 2017, pp. 1–6.
[38] Audio Engineering Society, "AES69-2015: AES standard for file exchange -Spatial acoustic data file format," 2015.
[39] A. Andreopoulou and B. F. G. Katz, "Identification of perceptually relevant methods of inter-aural time difference estimation," *J. Acoust. Soc. Amer.*, vol. 142, no. 2, pp. 588–598, 2017.
[40] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *J. Acoust. Soc. Amer.*, vol. 136, no. 2, pp. 791–802, 2014.
[41] T. May, S. Van De Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 1–13, Jan. 2011.
[42] P. Søndergaard and P. Majdak, "The auditory modeling toolbox," in *The Technology of Binaural Listening*, J. Blauert, Ed. Berlin Germany: Springer-Verlag, 2013, pp. 33–56.
[43] C. G. Buchanan and M. J. Newton, "Dynamic balanced model truncation of the spherical transfer function for use in structural HRTF models," in *Proc. Audio Eng. Soc. Conf. Audio Virtual Augmented Reality*, 2018, pp. 1–10.
[44] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1648–1661, 2005.
[45] E. A. Macpherson and A. T. Sabin, "Binaural weighting of monaural spectral cues for sound localization," *J. Acoust. Soc. Amer.*, vol. 121, no. 6, pp. 3677–3688, 2007.
[46] R. Bomhardt, "Anthropometric Individualization of Head-Related Transfer Functions - Analysis and Modeling," Ph.D. Dissertation, Inst. Tech. Acoust., RWTH Aachen Univ., Aachen, Germany, 2017.
[47] A. B. Watson and D. G. Pelli, "QUEST: A Bayesian adaptive psychometric method," *Perception Psychophys.*, vol. 33, no. 2, pp. 113–120, 1983.
[48] ITU BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems," 2015, pp. 1–35.
[49] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A spatial audio quality inventory (SAQI)," *Acta Acustica United Acustica*, vol. 100, no. 5, pp. 984–994, 2014.
[50] N. A. Gumerov and R. Duraiswami, "Computation of scattering from N spheres using multipole reexpansion," *J. Acoust. Soc. Amer.*, vol. 112, no. 6, pp. 2688–2701, 2002.
[51] R. Adelman, N. A. Gumerov, and R. Duraiswami, "Semi-analytical computation of acoustic scattering by spheroids and disks," *J. Acoust. Soc. Amer.*, vol. 136, no. 6, pp. EL405–EL410, 2014.
[52] J. Fels, R. Bomhardt, and F. Pausch, "Investigation on localization performance using smoothed individual head-related transfer functions," in *Proc. 42nd DAGA*, 2016, pp. 86–88.
[53] H. Wallach, "On sound localization," *J. Acoust. Soc. Amer.*, vol. 10, no. 4, pp. 270–274, 1939.
[54] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of HRTFs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. IV45–IV48.
[55] D. N. Zotkin, R. Duraiswami, E. Grassi, and N. A. Gumerov, "Fast head-related transfer function measurement via reciprocity," *J. Acoust. Soc. Amer.*, vol. 120, no. 4, pp. 2202–2215, 2006.
[56] W. Zhang, T. D. Abhayapala, R. A. Kennedy, and R. Duraiswami, "Insights into head-related transfer function: Spatial dimensionality and continuous representation," *J. Acoust. Soc. Amer.*, vol. 127, no. 4, pp. 2347–2357, 2010.
[57] M. Pollow *et al.*, "Calculation of head-related transfer functions for arbitrary field points using spherical harmonics decomposition," *Acta Acustica United Acustica*, vol. 98, no. 1, pp. 72–82, 2012.

Christoph Pörschmann received the Diploma in electrical engineering from the Ruhr-Universität Bochum, Bochum, Germany, and Uppsala Universitet, Uppsala, Sweden, in 1995, and the Doctoral Degree (Dr.-Ing.) from the Faculty of Electrical Engineering and Information Technology, Ruhr-Universität Bochum in 2001 as a result of his research at the Institute of Communication Acoustics.

Since 2004, he is a Professor of Acoustics at TH Cologne - University of Applied Sciences. His research interests include virtual acoustics, spatial hearing and the related perceptual processes.

Dr. Pörschmann is member of the German Acoustical Society and the Acoustical Society of America.

Johannes M. Arend received the B.Eng. degree in media technology from HS Düsseldorf, Düsseldorf, Germany, in 2011 and the M.Sc. degree in media technology from TH Köln, Cologne, Germany, in 2014. Since 2015, he has been a Research Fellow and working toward the Ph.D. degree at TH Köln and TU Berlin in the field of spatial audio with a focus on binaural technology, auditory perception, and audio signal processing.

Fabian Brinkmann received the M.A. degree (magister artium) in communication sciences and technical acoustics in 2011 from TU Berlin, Berlin, Germany where he is currently working toward the Ph.D. degree in the field of signal processing and evaluation approaches for spatial audio.

Since 2011, he has been a Research Associate with the Audio Communication Group from TU Berlin and is associated to the DFG research consortium SEACEN.

Correspondence

Correction to "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling"

Christoph Pörschmann, Johannes M. Arend, and Fabian Brinkmann

In the paper "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling," published in IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 27, Issue: 6, June 2019) [1], Eq. (2) and Eq. (6) are incorrect.

Correction Eq. (2): For the definition of azimuth ϕ and elevation θ used throughout the paper, the argument of the associated Legendre functions P_n^m in Eq. (2) of the paper must be $\sin\theta$ and not $\cos\theta$, as mistakenly specified by us. The correct equation describing the spherical harmonics Y_n^m of order n and mode/degree m according to the used definition of ϕ and θ is given by

$$Y_n^m(\theta,\phi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(\sin\theta)\mathrm{e}^{\mathrm{i}m\phi}\,, \tag{1}$$

with $\mathrm{i} = \sqrt{-1}$ the imaginary unit.

Correction Eq. (6): This equation describes the sound field on an open sphere and not the sound field on a rigid sphere as explained in the paper. The correct equation describing the sound field on a rigid sphere for unite amplitude incident plane waves and the corresponding rigid sphere transfer functions (STF) is given by [2] [3, Eqs. (2.43), (2.61)]

$$H_{STF}(\omega,\Omega_g) = \sum_{n=0}^{N_{high}} \sum_{m=-n}^{n} 4\pi\mathrm{i}^n \left[j_n(kr) - \frac{j_n'(kr)}{h_n^{(2)'}(kr)} h_n^{(2)}(kr) \right] * Y_n^m(\Omega_e) Y_n^m(\Omega_g)^*, \tag{2}$$

with j_n the spherical Bessel function of the first kind, $h_n^{(2)}$ the spherical Hankel function of the second kind, and j_n' and $h_n^{(2)'}$ their derivatives.[1] The imaginary unit is defined as $\mathrm{i} = \sqrt{-1}$, r denotes the radius, $k = \frac{\omega}{c}$ with ω the angular frequency, and c the speed of sound. Y_n^m denotes the complex spherical harmonics functions of order n and mode/degree m, Ω_e the ear position at $\phi = \pm 90°$ and $\theta = 0°$, Ω_g the incidence directions of the plane waves and the notation $(\cdot)^*$ denotes complex conjugation.

Comment: In the Matlab-based implementation of the SUpDEq-method, which is available on https://github.com/AudioGroupCologne/SUpDEq, the correct spherical harmonics equation according to (1) above as well as the correct rigid sphere transfer functions according to (2) above have been implemented. Thus, the correct spatial Fourier transform and correct rigid sphere transfer functions were applied for the processing described in the paper. The results presented in the paper are therefore still valid without any restriction.

Acknowledgment

The authors would like to thank Dr. Jens Ahrens (Chalmers University of Technology, Division of Applied Acoustics) for pointing out and discussing the error.

References

[1] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Directional equalization of sparse head-related transfer function sets for spatial upsampling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 6, pp. 1060–1071, Jun. 2019.

[2] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *J. Acoust. Soc. Amer.*, vol. 116, no. 4, pp. 2149–2157, 2004.

[3] B. Rafaely, *Fundamentals of Spherical Array Processing*. Berlin, Germany: Springer-Verlag, 2015.

[4] V. Tourbabin and B. Rafaely, "On the consistent use of space and time conventions in array processing," *Acta Acustica United Acustica*, vol. 101, no. 3, pp. 470–473, 2015.

Manuscript received May 3, 2020; revised July 7, 2020; accepted July 15, 2020. Date of publication July 21, 2020; date of current version August 3, 2020. This work was supported in part by the German Federal Ministry of Education and Research (BMBF 03FH014IX5-NarDasS) and in part by the German Research Foundation (DFG WE 4057/3-2). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Federico Fontana. *(C. Pörschmann and J. M. Arend contributed equally to this work.) (Corresponding author: Christoph Pörschmann.)*

Christoph Pörschmann is with the Institute of Communications Engineering, TH Köln - University of Applied Sciences, D-50679 Cologne, Germany (e-mail: christoph.poerschmann@th-koeln.de).

Johannes M. Arend is with the Institute of Communications Engineering, TH Köln - University of Applied Sciences, D-50679 Cologne, Germany, and also with the Audio Communication Group, Technical University of Berlin, D-10587 Berlin, Germany (e-mail: johannes.arend@th-koeln.de).

Fabian Brinkmann is with the Audio Communication Group, Technical University of Berlin, D-10587 Berlin, Germany (e-mail: fabian.brinkmann@tu-berlin.de).

Digital Object Identifier 10.1109/TASLP.2020.3010608

[1] Please note the dependency of Eq. (2) on the Fourier transform kernel [4, Table I]. We used $p(\omega) = \int_{-\infty}^{\infty} p(t) e^{-\mathrm{i}\omega t} \mathrm{d}t$ as the Fourier transform of the pressure signal $p(t)$.

2.6 Obtaining Dense HRTF Sets from Sparse Measurements in Reverberant Environments

Pörschmann, C., & Arend, J. M. (2019). In *Proc. of the AES International Conference on Immersive and Interactive Audio (IIA), York, UK* (pp. 1–10).

Pörschmann, C., & Arend, J. M. (2020). Correction to "Obtaining Dense HRTF Sets from Sparse Measurements in Reverberant Environments."

Audio Engineering Society

Conference Paper 15

Presented at the Conference on
Immersive and Interactive Audio
2019 March 27 – 29, York, UK

Obtaining Dense HRTF Sets from Sparse Measurements in Reverberant Environments

Christoph Pörschmann[1] and Johannes M. Arend[1,2]

[1] *Institute of Communications Engineering, TH Köln - University of Applied Sciences, D-50679 Cologne, Germany*
[2] *Audio Communication Group, Technical University Berlin, D-10587 Berlin, Germany*

Correspondence should be addressed to Christoph Pörschmann (christoph.poerschmann@th-koeln.de)

ABSTRACT

The paper describes a method for obtaining spherical sets of head-related transfer functions (HRTFs) based on a small number of measurements in reverberant environments. For spatial upsampling, we apply HRTF interpolation in the spherical harmonics (SH) domain. However, the number of measured directions limits the maximal accessible SH order, resulting in order-limitation errors and a restricted spatial resolution. Thus, we propose a method which reduces these errors by a directional equalization based on a spherical head model prior to the SH transform. To enhance the valid range of a subsequent low-frequency extension towards higher frequencies, we perform the extension on the equalized dataset. Finally, we apply windowing to the impulse responses to eliminate room reflections from the measured HRTF set. The analysis shows that the method for spatial upsampling influences the resulting HRTF sets more than degradations due to room reflections or due to distortions of the loudspeakers.

1 Introduction

A spatial presentation of sound sources is a fundamental element of virtual acoustic environments (VAEs). For this, monaural and binaural cues, which are mainly caused by the shape of the pinna and the head, need to be considered. While spectral information is the main cue to determine elevation, we use differences between the signals reaching the left and the right ear for lateral localization. These binaural differences manifest in interaural time differences (ITDs) and interaural level differences (ILDs). In many headphone-based VAEs, head-related transfer functions (HRTFs) are applied to describe the sound incidence from a source, which is typically in the far-field, to the left and right ear incorporating both, monaural and the binaural cues [1]. To adequately capture these cues for all directions of incidence, a high number of HRTFs is required. Several authors suggested to describe complete sets of HRTFs in spherical harmonics (SH) domain [2, 3]. Here, the HRTF set, measured on a spherical grid, is decomposed into spherical base functions of different orders n, where higher orders correspond to a higher spatial resolution. The use of sparse HRTF sets results in a limited order in the SH domain and involves an incomplete description of the spatial properties of the HRTF set. To completely consider these properties and to avoid spatial aliasing, an order $N \geq kr$ with $k = \omega/c$, and r being the head radius is required [4, 5]. Assuming $r = 8.75\,\text{cm}$ and $c = 343\,\text{m/s}$ leads to $N = 32$ for performing a nearly perfect interpolation of HRTFs for frequencies up to 20 kHz requiring at least $G = 1089$

measured directions. Different studies examined artifacts of order-limited HRTF representations, with the aim to determine the optimal number of required measurement positions. Bernschütz [5] proposed a spatial resampling of the HRTFs at the desired low order. By this, high order energy of the HRTFs is mirrored to lower orders, and the low-pass effect of order truncation is reduced. Ben-Hur et al. [6] suggested a post-hoc equalization of measured HRTF data, which reduces artifacts caused by an order truncation. Recently Alon et al. [7] proposed to minimize the aliasing errors by incorporating statistics calculated from a set of reference HRTFs. Other studies [8, 9] examined to what extent time aligning the HRTFs (by eliminating their linear phase components) before performing the SH transform affects the required SH order, suggesting that this efficiently reduces artifacts of order limitation.

In a recent paper we introduced the SUpDEq (Spatial Upsampling by Directional Equalization) method [10], which considers other direction-dependent temporal and spectral components as well. The method removes frequency-dependent ITDs and ILDs as well as elevation-dependent spectral features from the HRTFs. For this we apply a spectral division (equalization) of the HRTF with a corresponding equalization function prior to the SH transform. We used a directional rigid sphere transfer function (STF) which can be regarded as a simplified HRTF set as equalization function. After spatial upsampling by SH interpolation, a de-equalization by means of a spectral multiplication with the same equalization function is performed to recover a spatially upsampled HRTF set.

In this paper we use the SUpDEq method to enhance HRTF sets which were measured in a moderately reverberant acoustic environment. In addition to the reduced number of required directions for the acquisition of a complete HRTF set, we analyze further items for which the directional equalization can be beneficial. We investigate if a low-frequency extension, which is typically used to eliminate inaccuracies of the measurements at low frequencies, can be directly applied to the spatially equalized dataset. Furthermore, we examine to what extent the influence of early reflections and reverberance of the room can be eliminated from the measurement data by appropriate windowing of the measured head-related impulse responses (HRIRs). The paper is structured as follows: In Section 2 we describe the SUpDEq method, its combination with the low-frequency extension, and the different steps of enhancing sparse HRTF sets measured in reverberant environments. In Section 3 we evaluate the approach based on a comparison of measurements in an anechoic and a reverberant environment and analyze the influences of using a low-cost loudspeakers for the measurement.

2 Methods

2.1 Spatial Upsampling by Directional Equalization (SUpDEq)

A spherical dataset $H(\omega,\Omega_g)$ can be described in the SH domain by the SH coefficients $f_{nm}(\omega)$ that are obtained via the SH transform, often also referred to as spatial (or spherical) Fourier transform [2, 3]

$$f_{nm}(\omega) = \sum_{g=1}^{G} H(\omega,\Omega_g) Y_n^m(\Omega_g)^* \beta_g \,, \qquad (1)$$

with ω the temporal frequency in radians, β_g the sampling weights, and the G discrete HRTF-angles $\Omega_g = \{(\phi_1,\theta_1),\ldots,(\phi_G,\theta_G)\}$ at azimuth ϕ, and elevation θ. The complex conjugate is given by $(\cdot)^*$, Y_n^m denotes the SH of order n and mode m.

$$Y_n^m(\theta,\phi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(cos\theta) e^{im\phi} \,, \qquad (2)$$

with the associated Legendre functions P_n^m, $i=\sqrt{-1}$ the imaginary unit. The inverse spatial Fourier transform can be used to recover H at arbitrary angles

$$\widehat{H}(\omega,\Omega) = \sum_{n=0}^{N} \sum_{m=-n}^{n} f_{nm}(\omega) Y_n^m(\Omega) \,, \qquad (3)$$

where N denotes the maximal order. If H is strictly order-limited, a sufficient choice of N results in $H=\widehat{H}$. Depending on the spatial sampling grid Ω_g, the coefficients f_{nm} can be calculated up to a maximum order N. The number of measured directions G directly corresponds to the maximum order N by $G \propto (N+1)^2$. In case the order of H exceeds N, spatial aliasing occurs [3]. In this context, an appropriate preprocessing that reduces the spatial complexity of H will directly relax the requirement on G.

The SUpDEq method performs spatial upsampling of a sparse HRTF set as shown in Fig. 1. Please note that the entire processing is identical for the left and right ear signals, and corresponding subscripts were omitted

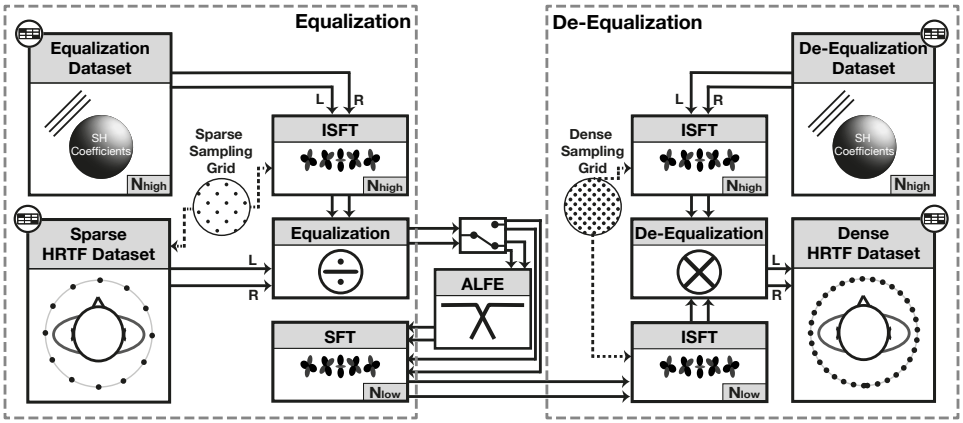


Fig. 1: Block diagram of the SupDEq method. Left panel: A sparse HRTF set is equalized on the corresponding sparse sampling grid. The set is then transformed to SH domain with $N = N_{low}$. Mid panel: Prior to the SH transform optionally a low-frequency extension can be performed on the equalized dataset. Right panel: The equalized set is de-equalized on a dense sampling grid, resulting in a dense HRTF set.

in the following for ease of display. In a first step, the sparse HRTF set H_{HRTF} measured at S sampling points $\Omega_s = \{(\phi_1, \theta_1), \ldots, (\phi_S, \theta_S)\}$ is equalized with an appropriate equalization dataset H_{EQ}

$$H_{HRTF,EQ}(\omega, \Omega_s) = \frac{H_{HRTF}(\omega, \Omega_s)}{H_{EQ}(\omega, \Omega_s)}. \tag{4}$$

The equalization dataset is intended to remove the directional dependency in H_{HRTF} to a certain degree with the goal to reduce the required order for the SH transform. Different equalization datasets can be applied – throughout this study an analytic rigid sphere transfer function (STF) is used [2].

$$H_{STF}(\omega, \Omega_g) = P 4\pi \sum_{n=0}^{N_{high}} \sum_{m=-n}^{n} i^n j_n(kr) Y_n^m(\Omega_e) Y_n^m(\Omega_g)^* \tag{5}$$

with $i = \sqrt{-1}$ the imaginary unit, j_n the spherical Bessel function of the first kind, and the ear position Ω_e at $\phi = 90°$ and $\theta = 0°$. P denotes an arbitrary sound pressure.

The STF can thus be regarded as a simplified HRTF set which features basic temporal and spectral components but does not carry information on the shape of the outer ears nor the fine structure of the head. Due to this, the equalization indeed considerably reduces the spatial order of H_{HRTF}, and thus also the spatial complexity, and the required number of sampling points S. Except for an angular shift in azimuth, the set is identical for the left and the right ear. As the equalization dataset is based on an analytic description, it can be determined at a freely chosen maximal order, typically, a high order $N_{high} \geq 35$.

SH coefficients $f_{EQ,nm}$ for the equalized sparse HRTF set are obtained in the second step using Eq. 1 with $H_{HRTF,EQ}(\omega, \Omega_s)$ up to an appropriate low maximal order N_{low}. In a third step, an upsampled HRTF set $\widehat{H}_{HRTF,EQ}$ is calculated for a dense sampling grid $\Omega_d = \{(\phi_1, \theta_1), \ldots, (\phi_D, \theta_D)\}$, with $D \gg S$ by using Eq. 3. In a fourth and final step, HRTFs are reconstructed by a subsequent de-equalization by means of spectral multiplication with a de-equalization dataset

$$\widehat{H}_{HRTF,DEQ}(\omega, \Omega_d) = \widehat{H}_{HRTF,EQ}(\omega, \Omega_d) H_{DEQ}(\omega, \Omega_d). \tag{6}$$

By this, energy which was transformed to lower orders in the first step is recovered in higher spatial orders. For de-equalization, again the STF as given in Eq. 5 is used. $H_{HRTF} = \widehat{H}_{HRTF,DEQ}$ holds if N_{low} and N_{high} are chosen appropriately. Otherwise, deviations will be caused by signal energy which, after the equalization, still is apparent at high modal orders $N > N_{low}$. Due to spatial aliasing, this signal energy is irreversibly mirrored to lower orders $N \leq N_{low}$ as described in [5], and we obtain $H_{HRTF} \approx \widehat{H}_{HRTF,DEQ}$. In Section 3.2 we analyze the influence of the order N_{low} and the benefit of the SUpDEq method compared to common (order-limited) SH interpolation.

2.2 Adaptive Low-Frequency extension (ALFE)

Typical measurements of HRTFs involve several limitations. First of all, in order to replicate a point source at low distances often small loudspeakers are used, which mostly fail to reproduce low frequencies at adequate sound pressure levels. This leads to HRTFs with a distinct low-frequency roll-off and needs to be compensated when the HRTFs are used for auralization.

Another relevant problem is the sound field of the measuring room. Even in an anechoic chamber, room modes and reflections arise below its cut-off frequency. If the measurements take place in reverberant environments, this impact increases towards higher frequencies as well. Distinct room reflections and resonances can be observed that strongly influence the measuring data. Because of this modal behavior, raw HRTF measurements show room and position dependent peaks and dips. To remove influences of reflections and resonances, windowing can be applied. However, depending on the window size, low-frequency components of the direct sound are often affected by the windowing as well. Thus, a reconstruction of the low-frequency components based on an appropriate model is required.

A well-suited approach is to replace the low-frequency range of the HRTFs by an analytic expression. In [11] an approach is described which assumes that for low frequencies (e.g. below 200 Hz), pinna and ear canal hardly affect the HRTF and even the shape of the head only has minor influence on the sound field. Accordingly it is reasonable to extend HRTFs towards lower frequencies and by this obtain a flat frequency response below a certain corner frequency. The basic idea of this so-called ALFE-algorithm is to attach a matched low-frequency extension, substituting the original low-frequency component. In this study we use a low-frequency extension in the frequency domain according to [12]. The extension applies linear cross-fading between the low-frequency component and the raw HRTF in a certain crossover frequency range (e.g. 200 Hz – 400 Hz). The level is calculated from the mean absolute values, while the phase is linearly extrapolated in the crossover frequency range.

2.3 Combination of SUpDEq and ALFE

Influences of diffraction and interferences from the HRTF set which are caused by a spherical head shape are removed by the spatial equalization (see Eq. 4). The remaining effects of the pinna shape and ear canal become relevant at much higher frequencies where the wavelength is in the range of their dimensions. Thus, when applying ALFE to the spatially equalized data set, the crossover frequency range can be set significantly higher. Subsequently, we integrated ALFE in SUpDEq to be applied optionally after the equalization (see Fig. 1). In our study we chose a crossover frequency range from 500 Hz to 750 Hz. Applying ALFE, the magnitude response of the equalized HRTF set becomes flat below this frequency range. Furthermore, we put specific care on the level adjustment of the extended components and set the level to a fixed direction-independent value for frequencies below the crossover frequency range. By this, for very low frequencies the de-equalized HRTF set equals the rigid STF which serves as an appropriate low-frequency HRTF model.

3 Technical Evaluation

3.1 HRTF Measurements

We performed HRTF measurements of a Neumann KU100 dummy head in anechoic and in reverberant environments. A Genelec 1029A loudspeaker, which has a flat on-axis frequency response from 50 Hz to 20 kHz (±3 dB) was used in this study. Furthermore, as an example of a low-cost loudspeaker, we tested an active battery-driven JBL Clip+ speaker. The speaker has a maximum power of 3.6 W and comprises a 40 mm transducer allowing a sound radiation in a frequency range from 160 Hz – 20 kHz. We used the VariSphear measurement system [13] for precise positioning of the dummy head at the spatial sampling positions and for capturing the HRTFs. The HRTFs were measured on a Lebedev full spherical grid with 2702 points. The excitation signal for all measurements was an emphasized sine sweep with +20 dB low shelf at 100 Hz (2^{18} samples at 48 kHz sampling rate, length 5.5 s). An RME Babyface audio interface was used as AD / DA converter and microphone preamp. For further details on the set-up and procedure please refer to [11] or [14].

For the reverberant environment, the measurements were done in a seminar room at TH Köln (see Fig. 2). The room has a reverberation time $T_{60}(500/1000Hz) = 0.7s$ and its shape approaches a shoebox and sizes 13 m × 7 m × 3 m (W×D×H). While most of the walls are highly reflective, the ceiling is covered with sound-absorbing material. The

Fig. 2: Measurement of the HRTF set in the seminar room at TH Köln using a Neumann KU100 artificial head mounted on the VariSphear device.

VariSphear was placed off-axis but near the center of the room. To compare the HRTF sets to a reference, we repeated the measurements with the Genelec speaker in the anechoic chamber at TH Köln. The chamber has dimensions of 4.5 m × 11.7 m × 2.3 m and a low cut-off frequency of about 200 Hz. In both rooms, sets of HRTFs were captured at 2 m distance. The height of the loudspeakers and of the dummy head was at 1.25 m and the acoustic center of the loudspeaker was always set to the ear level of the dummy head. Thus, in the reverberant environment the distance difference between the direct sound and the first reflection is about 1.2 m. Additionally, we captured for each set omnidirectional impulse responses at the acoustic center of the dummy head with a Microtech Gefell M296S microphone. These measurements provided the basis for the magnitude and phase compensation of the loudspeaker.

In a subsequent postprocessing, the raw measurement data were carefully truncated and windowed. Then we compensated the influence of the loudspeaker by inverse FIR filtering with the measured omnidirectional impulse response. The final length of each HRIR is 128 samples at a sampling rate of 48 kHz. The postprocessing is based on the implementation and description from [11] and [14]. In contrast to these studies, no ALFE-processing was done for the HRTF sets captured in the seminar room. For the reference datasets measured in the anechoic chamber we applied ALFE with a crossover frequency range from 200 Hz – 400 Hz.

3.2 SH order

We applied the SUpDEq method according to Section 2.1 for a radius of $r = 9.19$ cm which we calculated according to Algazi et al. [15] based on the dimensions of the Neumann KU100 dummy head. Furthermore, we included the ALFE processing as described in Section 2.2. In the following we compare the high density reference HRTFs ($N_{high} = 35$) to sparse HRTF sets processed with SUpDEq and to strictly order-limited HRTFs, obtained by means of SH interpolation without pre- or postprocessing. We obtained the sparse sets by spatially downsampling the dense sets on a Lebedev grid Ω_s ($N_{low} = 4$, 7, 10, 13; corresponding to 38, 86, 170, 266 sampling points).

First we analyze some exemplary directions for $N_{low} = 4$. We extracted left ear HRTFs for the frontal ($\phi = 0°, \theta = 0°$) and the more critical contralateral direction ($\phi = 270°, \theta = 0°$). Fig. 3 illustrates the respective magnitude and impulse responses. We observed great similarities between the de-equalized HRTF and the reference for the frontal direction, apart from a slight ripple above approximately 2 kHz. In contrast, the order-limited HRTF clearly suffers from the typical high-frequency deviations for the frontal direction due to spatial aliasing (see e.g. [5, 6, 8]). For contralateral sound incidence, the magnitude responses of both HRTFs show increased distortions above 1 kHz. For localization in the horizontal plane, the temporal structure is even more important. Here, for the frontal and contralateral directions, the de-equalized HRIR follows the reference. In contrast, the strictly order-limited HRIR shows distinct pre-ringing artifacts, especially for the contralateral incidence. Informal tests on our measured data showed that the artifacts decrease with increasing order but remain visible even at $N_{low} = 15$. Generally, contralateral directions are much more critical than frontal or ipsilateral directions, mostly because diffraction of the head influences sound incidence at the contralateral ear and is not completely matched by the applied (de-)equalization dataset.

To analyze the deviations to a reference set over all T measured directions $\Omega_t = \{(\phi_1, \theta_1), \ldots, (\phi_T, \theta_T)\}$ of the sampling grid, we calculated the spectral differences averaged across all 2702 measured directions:

$$\Delta G_f(\omega) = \frac{1}{N_{\Omega_t}} \sum_{\Omega_t} |20lg \frac{|H_{HRTF,REF}(\omega, \Omega_t)|}{|H_{HRTF,TEST}(\omega, \Omega_t)|}|, \quad (7)$$

with $H_{HRTF,REF}$ describing the reference HRTF set and $H_{HRTF,TEST}$ the respectively processed sparse HRTF set. Figure 4 illustrates the frequency-dependent spectral differences $\Delta G_f(\omega)$ at $N_{low} = 4, 7, 10, 13$ for the SUpDEq method and for an strictly order-limited interpolation. The analysis refers to the dataset measured

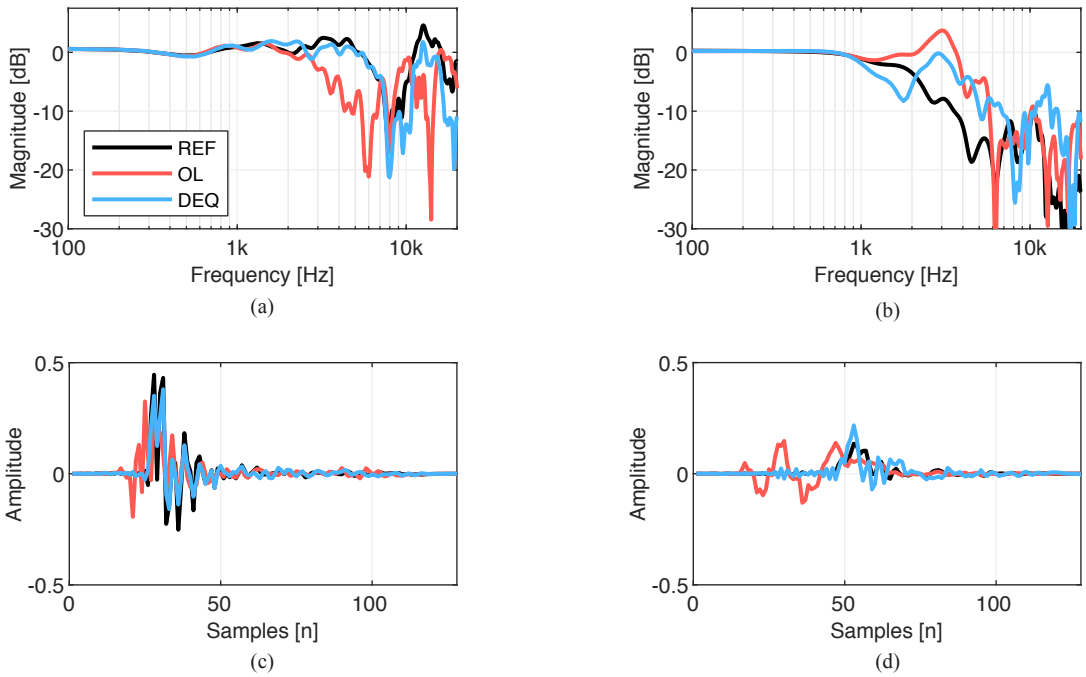


Fig. 3: Left ear magnitude (top) and impulse (bottom) responses, extracted from the order-limited set (red) and the SUpDEq-processed set (blue) by SH interpolation based on a sparse HRTF set with 38 sampling points ($N_{low} = 4$). The magnitudes and the impulse responses from the reference ($N_{high} = 35$) are shown for comparison (black). (a), (c) Front direction ($\phi = 0°, \theta = 0°$). (b), (d) Contralateral direction ($\phi = 270°, \theta = 0°$). All measurements were carried out in the anechoic chamber with the Genelec speaker.

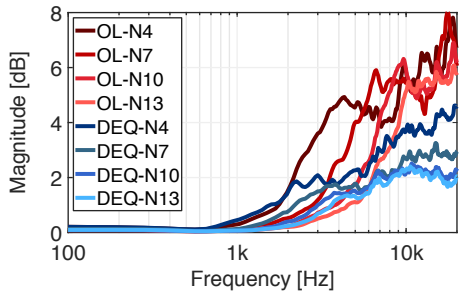


Fig. 4: Spectral differences $\Delta G_f(\omega)$ in dB (left ear) between reference HRTF set ($N_{high} = 35$) and strictly order-limited sets (red) and SUpDEq-processed sets (blue) for $N_{low} = 4, 7, 10, 13$.

in the anechoic chamber. Two things can be easily observed. First, the spectral differences are significantly smaller for the SUpDEq method than for order-limited interpolation. Second, for order-limited interpolation the spectral differences increase distinctively above 2 dB at aliasing frequencies between 2 and 6 kHz, depending on N. For the SUpDEq method, the spectral differences are generally lower and show a much more gentle rise. Here, the differences exceed 2 dB at frequencies between 3 and 9 kHz and remain between 2 and 4 dB, even above 10 kHz.

In a next step we analyzed the spatial distribution of the differences and calculated the directional deviation across all frequencies as

$$\Delta G_{sp}(\Omega_t) = \frac{1}{N_\omega} \sum_{\omega} \left| 20lg \frac{|\, H_{HRTF,REF}(\omega, \Omega_t)\,|}{|\, H_{HRTF,TEST}(\omega, \Omega_t)\,|} \right|, \quad (8)$$

Analyzing this data for different orders N_{low} we found that the order-limited interpolation results in distinct spectral differences spread over the entire angular range. In contrast, the SUpDEq method leads to differences mainly located at contralateral directions. For $N_{low} = 4$ we determined a maximum of $\Delta G_{sp,max} = 11\,dB$ for the order-limited interpolation at $\phi = 265°$ and $\theta = -13°$. The SUpDEq method results in an only slightly smaller maximal spectral difference of $\Delta G_{sp,max} = 9\,dB$ at

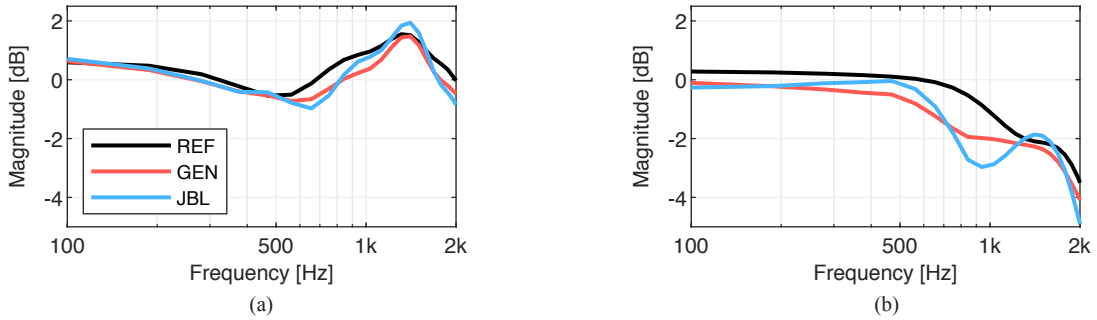


Fig. 5: Influence of the low-frequency extension performed on the equalized dataset. HRTFs for the Genelec (red) and the JBL loudspeaker (blue) measured in the seminar room and for comparison the reference (black). Left ear magnitude for (a) frontal incidence ($\phi = 0°, \theta = 0°$) and (b) contralateral incidence ($\phi = 270°, \theta = 0°$).

$\phi = 273°$ and $\theta = +4°$. Further details on the spatial distribution of the differences are described in [10].

3.3 Influence of Low-frequency Extension

In the following section we consider full density HRTF sets with $N_{high} = 35$ and analyze the influence of the low-frequency extension only. The extension is performed on the equalized data of the HRTF sets measured in the seminar room with a crossover frequency range from 500 Hz – 750 Hz according to Section 2.3. This means that for frequency components below 500 Hz, influences of reflections and resonances as well as a band-limitation of the loudspeaker are not relevant.

Figure 5 shows results for frontal ($\phi = 0°, \theta = 0°$) and for contralateral sound incidence ($\phi = 270°, \theta = 0°$) for both speakers. The frequency responses are not completely flat below the crossover frequency range as they follow the frequency response of a rigid sphere. This is especially for the frontal sound incidence beneficial, where variations of the magnitude response below 500 Hz are still relevant and above 1 dB. The results are in good agreement to the reference, which was processed with ALFE on the non-equalized data (crossover frequency range 200 Hz – 400 Hz). The deviations between reference and the datasets with ALFE integrated in SUpDEq are relatively small for both speakers. Below the crossover frequency range the average difference $\Delta G_f(\omega)$ is far below 1 dB (see Fig. 7).

Thus, performing the extension on the equalized dataset appropriately considers a spherical head geometry in the low frequency range. This seems to be adequate to significantly shift the range of crossover frequencies upwards.

3.4 Room and loudspeaker influence

As described above, both the acoustic conditions of the measuring room and the frequency response of the loudspeaker have strong influence on the raw data. The first reflection is the floor reflection and reaches the ear at a time delay of 3.5 ms corresponding to 168 samples at 48 kHz sampling rate. Thus, in order to eliminate room reflections and to correct the frequency response, we applied postprocessing and windowed the HRIRs to a length of 128 samples as described in Section 3.1. As above, we consider full density sets with $N_{high} = 35$. For low-frequency components, we applied ALFE in the same way as in 3.3. Now we compare the post-processed sets measured with the two loudspeakers in the seminar room to a reference measurement made in the anechoic chamber.

As shown in Fig. 6 for the two directions ($\phi = 0°, \theta = 0°$ and $\phi = 270°, \theta = 0°$) no influence of early reflections can be observed in the impulse responses. Both for the Genelec and the JBL speaker, the HRTFs and the HRIRs are very similar and match the reference data quite well. To generalize these exemplary results, we calculated the average deviations $\Delta G_f(\omega)$ between the reference dataset and the measurements in the seminar room (Fig. 7). For the Genelec speaker these differences are below 1.5 dB for frequencies up to 10 kHz and might be induced by the remaining influence of the measuring room. For the JBL speaker the differences are larger, but still below 3 dB for frequencies up to 12 kHz. Thus the influence of the low-cost loudspeaker is relevant, but is in a range which might be acceptable for many VR applications. Even the peak

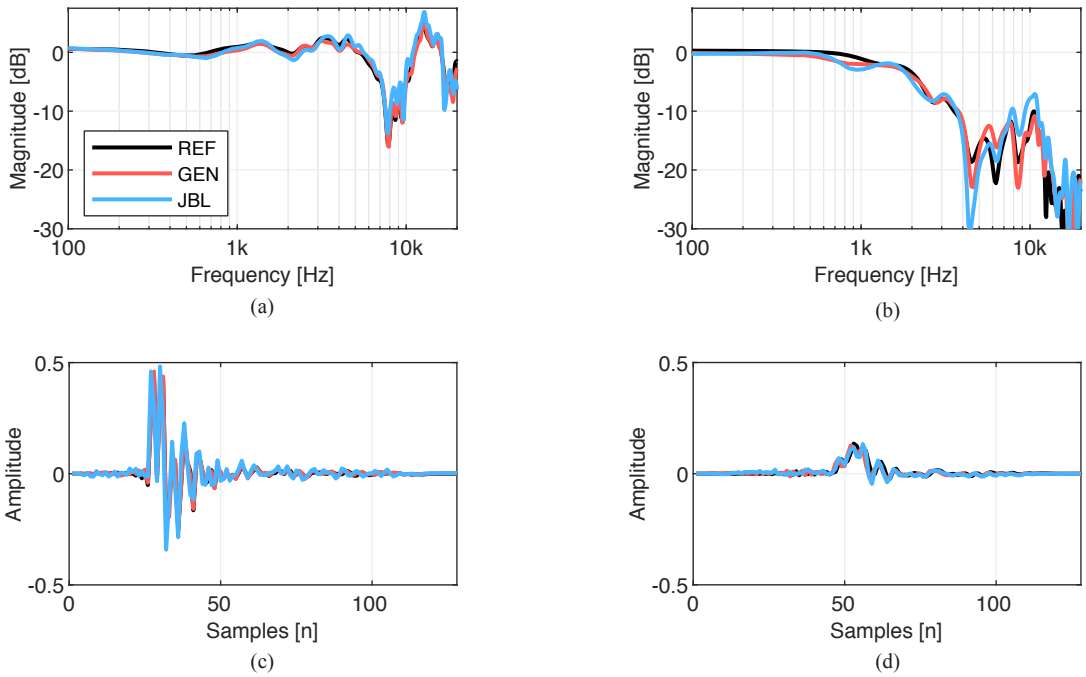


Fig. 6: Left ear magnitude (top) and impulse responses (bottom), extracted from the SUpDEq-processed set measured in the reverberant environment of the seminar room at $N_{high} = 35$ for the Genelec (red) and the JBL speaker (blue) compared to the reference measured with the Genelec speaker in the anechoic chamber (black). (a), (c) Front direction ($\phi = 0°, \theta = 0°$). (b), (d) Contralateral direction ($\phi = 270°, \theta = 0°$).

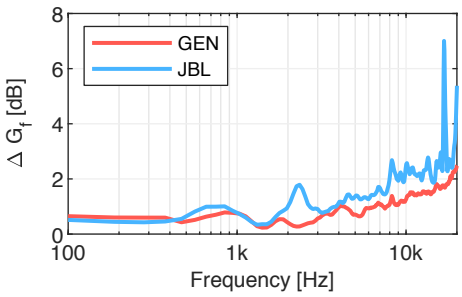


Fig. 7: Frequency-dependent spectral differences $\Delta G_f(\omega)$ between the SUpDEq-processed measurements in the seminar room for the Genelec and the JBL speaker and the reference measurement from the anechoic chamber.

with deviations of about 7 dB observed for the JBL speaker at about 16 kHz, which probably results from the speaker's strong linear distortions in this frequency region, might be uncritical for many applications. However, these deviations can as well be influenced by other factors, caused, e.g. by small variations in the measuring conditions and setup. Thus, the influence of the measuring room might even be lower than calculated from our measured data. Spatially the deviations are nearly equally distributed for both speakers in the lower hemisphere of the contralateral side and might be caused by influences of a floor reflection. We observed maximal values of $\Delta G_{sp,max} = 4.6\,dB$ at $\phi = 305°$ and $\theta = -10°$ for the Genelec and $\Delta G_{sp,max} = 5.1\,dB$ at $\phi = 269°$ and $\theta = -9°$ for the JBL speaker.

3.5 Combination of all influencing factors

Finally, we analyze how the combination of all degrading factors affects the measured and postprocessed HRTF set. These are the influence of the sparse HRTF set, of the room and of the low-cost loudspeaker. Generally, the temporal structure of the impulse responses is mainly unaffected by the processing steps (see Fig. 8). The influences of the contributing factors on the spectrum are shown in Figure 8 as well and are comparable to deviations due to the SH interpolation only. Figure 9 shows $\Delta G_f(\omega)$ for both speakers and various orders N_{low} for measurements in the seminar room. While

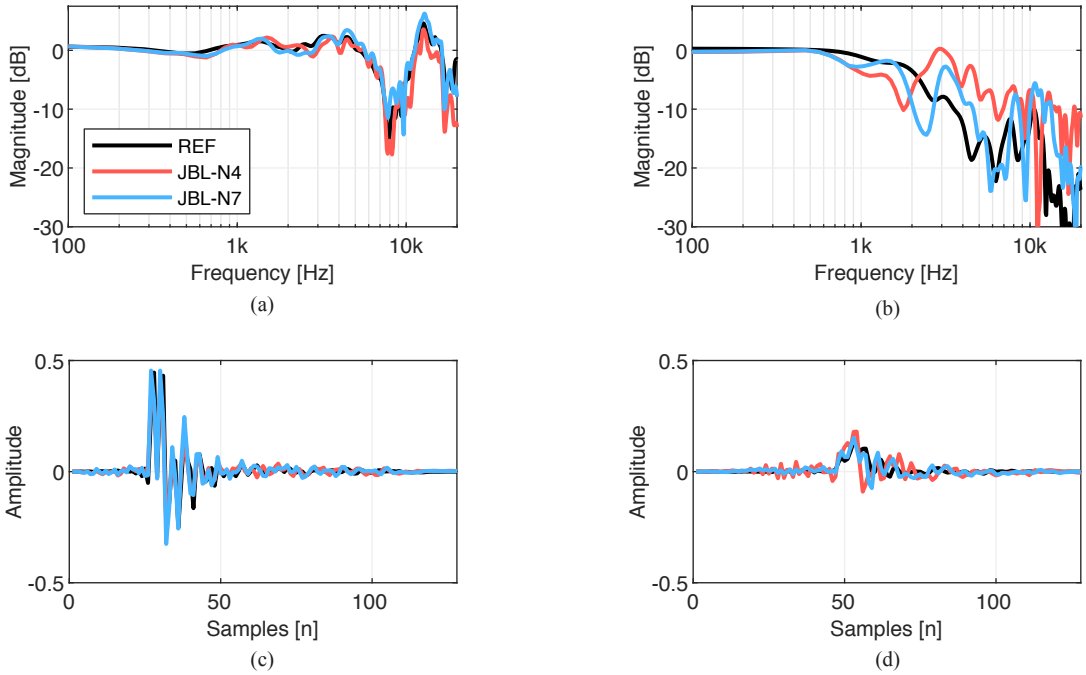


Fig. 8: Left ear magnitude (top) and impulse (bottom) responses, extracted from the SUpDEq-processed set measured in the reverberant environment of the seminar room with the JBL speaker for the orders $N_{low} = 4$ (red) and $N_{low} = 7$ (blue) compared to the reference set ($N_{high} = 35$) measured in the anechoic chamber (black). (a), (c) Front direction ($\phi = 0°, \theta = 0°$). (b), (d) Contralateral direction ($\phi = 270°, \theta = 0°$).

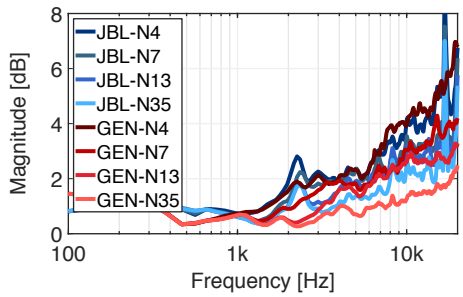


Fig. 9: Frequency-dependent spectral differences $\Delta G_f(\omega)$ of the SUpDEq-processed data from the seminar room for the Genelec and the JBL speaker for orders $N = 4, 7, 13, 35$ compared to the reference data from the anechoic chamber ($N_{high} = 35$).

for $N_{high} = 35$ the deviations are only caused by the different loudspeakers and rooms, for lower orders the influence of the spatial upsampling becomes more and more dominant and outweighs the deviations induced by the loudspeakers and the reverberant environment.

4 Conclusion

In this paper we compared sparse HRTF sets measured in a moderately reverberant room to reference measurements carried out in an anechoic chamber. We used the SUpDEq method for spatial upsampling of the sets and integrated a low-frequency extension into SUpDEq which basically attaches a frequency and phase response of a rigid sphere below a crossover frequency range between 500 – 750 Hz. Furthermore, to remove room reflections and resonances, we appropriately windowed the measured HRIRs.

The evaluation showed that upsampling sparse HRTF datasets by directional filtering as technically realized with the SUpDEq method is significantly better than a SH interpolation without any pre- or postprocessing, also referred to as strictly order-limited upsampling. However, low-frequency components need to be processed separately. For measurements in reverberant environments the crossover frequency between the low-frequency extension and the measured data has to be set so high that we recommend to perform this extension

on a directionally equalized dataset. By this, a spherical head geometry is approximated for low frequencies. For the remaining mid- and high-frequency components, appropriate windowing of the impulse responses in the time domain eliminates room reflections from the measured HRTF set. Finally, above the crossover frequency even strong linear distortions of low-cost speakers can be compensated in postprocessing by inverse filtering with a measured omnidirectional HRIR.

Generally speaking, according to our study a specific focus must be put on the spatial upsampling, more than on optimizing the room acoustics of the measuring room or on using specific high-quality loudspeakers for the measurements. Of course, our findings need to be verified for individually measured HRTF sets and be validated in a subsequent perceptual evaluation. The results of this study are of great relevance for applications where HRTFs need to be measured under non-optimal acoustic conditions. Even though the required number of directions of the sparse grid depends on the specific use case, we found that already from 38 HRTFs ($N_{low} = 4$) measured in a reverberant environment, a decent full-spherical dense HRTF set can be generated.

5 Acknowledgments

The research presented in this paper has been funded by the German Federal Ministry of Education and Research. Support Code: BMBF 03FH014IX5-NarDasS.

References

[1] Blauert, J., *Spatial Hearing - The Psychophysics of Human Sound Localization*, MIT Press, Cambridge, MA, revised edition, 1996.

[2] Williams, E. G., *Fourier Acoustics - Sound Radiation and Nearfield Acoustical Holography*, Academic Press, London, UK, 1999.

[3] Rafaely, B., *Fundamentals of Spherical Array Processing*, Springer-Verlag, Berlin Heidelberg, 2015.

[4] Rafaely, B., "Analysis and Design of Spherical Microphone Arrays," *IEEE Trans. on Speech and Audio Proc.*, 13(1), pp. 135–143, 2005.

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

[6] Ben-Hur, Z., Brinkmann, F., Sheaffer, J., Weinzierl, S., and Rafaely, B., "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *J. Acous. Soc. Am.*, 141(6), pp. 4087–4096, 2017.

[7] Alon, D. L., Ben-Hur, Z., Rafaely, B., and Mehra, R., "Sparse Head-Related Transfer Function Representation with Spatial Aliasing Cancellation," in *Proc. of the ICASSP 2018*, pp. 6792–6796, 2018.

[8] Zaunschirm, M., Schoerkhuber, C., and Hoeldrich, R., "Binaural rendering of Ambisonic signals by HRIR time alignment and a diffuseness constraint," *J. Acous. Soc. Am.*, 143(6), pp. 3616 – 3627, 2018.

[9] Brinkmann, F. and Weinzierl, S., "Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition," in *Proc. of the AES Int. Conf. on Audio for Virtual and Augmented Reality*, pp. 1–10, 2018.

[10] Pörschmann, C., Arend, J. M., and Brinkmann, F., "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling," *Manuscript, submitted for Publication*, 2018.

[11] Bernschütz, B., "A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100," in *Proc. of the 39th DAGA*, pp. 592–595, 2013.

[12] Xie, B., "On the low frequency characteristics of head-related transfer function," *Chinese J. Acoust.*, 28, pp. 1–13, 2009.

[13] Bernschütz, B., Pörschmann, C., Spors, S., and Weinzierl, S., "Entwurf und Aufbau eines variablen sphärischen Mikrofonarrays für Forschungsanwendungen in Raumakustik und Virtual Audio," in *Proc. of the 36th DAGA*, pp. 717–718, 2010.

[14] Arend, J. M., Neidhardt, A., and Pörschmann, C., "Measurement and Perceptual Evaluation of a Spherical Near-Field HRTF Set," in *Proc. of the 29th VDT International Convention*, pp. 52–55, 2016.

[15] Algazi, V., Avendano, C., and Duda, R. O., "Estimation of a Spherical-Head Model from Anthropometry," *J. Audio Eng. Soc.*, 49(6), pp. 472 – 479, 2001.

Obtaining Dense HRTF Sets from Sparse Measurements in Reverberant Environments

— Correction —

Christoph Pörschmann, Johannes M. Arend

Institute of Communications Engineering,
TH Köln - University of Applied Sciences

christoph.poerschmann@th-koeln.de
johannes.arend@th-koeln.de

Technology
Arts Sciences
TH Köln

Correction

In the paper "Obtaining Dense HRTF Sets from Sparse Measurements in Reverberant Environments", published in Proceedings of the AES International Conference on Immersive and Interactive Audio (IIA), York, UK, 2019 [1], Eq. (2) and Eq. (5) are incorrect.

Correction Eq. (2) For the definition of azimuth ϕ and elevation θ used throughout the paper, the argument of the associated Legendre functions P_n^m in Eq. (2) of the paper must be $\sin\theta$ and not $\cos\theta$, as mistakenly specified by us. The correct equation describing the spherical harmonics Y_n^m of order n and mode/degree m according to the used definition of ϕ and θ is given by

$$Y_n^m(\theta,\phi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(\sin\theta)\mathrm{e}^{\mathrm{i}m\phi}\,, \tag{1}$$

with $\mathrm{i} = \sqrt{-1}$ the imaginary unit.

Correction Eq. (5) This equation describes the sound field on an open sphere and not the sound field on a rigid sphere as explained in the paper. The correct equation describing the sound field on a rigid sphere for unite amplitude incident plane waves and the corresponding rigid sphere transfer functions (STF) is given by [2][3, Eqs. (2.43), (2.61)]

$$H_{STF}(\omega,\Omega_g) = \sum_{n=0}^{N_{high}} \sum_{m=-n}^{n} 4\pi\mathrm{i}^n \left[j_n(kr) - \frac{j_n'(kr)}{h_n^{(2)'}(kr)} h_n^{(2)}(kr) \right] Y_n^m(\Omega_e) Y_n^m(\Omega_g)^*, \tag{2}$$

with j_n the spherical Bessel function of the first kind, $h_n^{(2)}$ the spherical Hankel function of the second kind, and j_n' and $h_n^{(2)'}$ their derivatives.[1] The imaginary unit is defined as $\mathrm{i} = \sqrt{-1}$, r denotes the radius, $k = \frac{\omega}{c}$ with ω the angular frequency, and c the speed of sound. Y_n^m denotes the complex spherical harmonics functions of order n and mode/degree m, Ω_e the ear position at $\phi = \pm 90°$ and $\theta = 0°$, Ω_g the incidence directions of the plane waves and the notation $(\cdot)^*$ denotes complex conjugation.

Comment In the Matlab-based implementation of the SUpDEq-method, which is available on https://github.com/AudioGroupCologne/SUpDEq, the correct spherical harmonics equation according to Eq. (1) above as well as the correct rigid sphere transfer functions according to Eq. (2) above have been implemented. Thus, the correct spatial Fourier transform and correct rigid sphere transfer functions were applied for the processing described in the paper. The results presented in the paper are therefore still valid without any restriction.

Acknowledgment We would like to thank Dr. Jens Ahrens (Chalmers University of Technology, Division of Applied Acoustics) for pointing out and discussing the error.

References

[1] C. Pörschmann and J. M. Arend, "Obtaining Dense HRTF Sets from Sparse Measurements in Reverberant Environments," in *Proceedings of the AES International Conference on Immersive and Interactive Audio (IIA), York, UK*, 2019, pp. 1–10.

[2] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2149–2157, 2004.

[3] ——, *Fundamentals of Spherical Array Processing*. Berlin, Germany: Springer, 2015.

[4] V. Tourbabin and B. Rafaely, "On the Consistent Use of Space and Time Conventions in Array Processing," *Acta Acust. united Ac.*, vol. 101, no. 3, pp. 470–473, 2015.

[1] Please note the dependency of Eq. (2) on the Fourier transform kernel [4, Table I]. We used $p(\omega) = \int_{-\infty}^{\infty} p(t)e^{-\mathrm{i}\omega t}\mathrm{d}t$ as the Fourier transform of the pressure signal $p(t)$.

3 Parametric Spatial Audio

3.1 Six-Degrees-of-Freedom Parametric Spatial Audio Based on One Monaural Room Impulse Response

Arend, J. M., Amengual Garí, S. V., Schissler, C., Klein, F., & Robinson, P. W. (2021). *J. Audio Eng. Soc.*, *69*(7/8), 557–575. https://doi.org/10.17743/jaes.2021.0009

OPEN ACCESS Freely available online
J. M. Arend, S. V. A. 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, no. 7/8, pp. 557–575, (2021 July/August).
DOI: https://doi.org/10.17743/jaes.2021.0009

Six-Degrees-of-Freedom Parametric Spatial Audio Based on One Monaural Room Impulse Response

JOHANNES M. AREND,[1,2]* *AES Student Member*,
(johannes.arend@th-koeln.de)

SEBASTIÀ V. AMENGUAL GARÍ,[3] *AES Associate Member*, **CARL SCHISSLER,**[3]
(samengual@fb.com) (cschissler@fb.com)

FLORIAN KLEIN,[4]* **AND PHILIP W. ROBINSON,**[3] *AES Member*
(florian.klein@tu-ilmenau.de) (philrob22@fb.com)

[1]*Institute of Communications Engineering, TH Köln - University of Applied Sciences, Cologne, D-50679, Germany*
[2]*Audio Communication Group, Technical University of Berlin, Berlin, D-10587, Germany*
[3]*Facebook Reality Labs Research, Redmond, WA 98052, USA*
[4]*Electronic Media Technology Lab, Technical University of Ilmenau, Ilmenau, D-98693, Germany*

Parametric spatial audio rendering is a popular approach for low computing capacity applications, such as augmented reality systems. However most methods rely on spatial room impulse responses (SRIR) for sound field rendering with 3 degrees of freedom (DoF), i.e., for arbitrary head orientations of the listener, and often require multiple SRIRs for 6-DoF rendering, i.e., when additionally considering listener translations. This paper presents a method for parametric spatial audio rendering with 6 DoF based on one monaural room impulse response (RIR). The scalable and perceptually motivated encoding results in a parametric description of the spatial sound field for any listener's head orientation or position in space. These parameters form the basis for the binaural room impulse responses (BRIR) synthesis algorithm presented in this paper. The physical evaluation revealed good performance, with differences to reference measurements at most tested positions in a room below the just-noticeable differences of various acoustic parameters. The paper further describes the implementation of a 6-DoF real-time virtual acoustic environment (VAE) using the synthesized BRIRs. A pilot study assessing the plausibility of the 6-DoF VAE showed that the system can provide a plausible binaural reproduction, but it also revealed challenges of 6-DoF rendering requiring further research.

0 INTRODUCTION

Augmented reality (AR) applications must provide perceptually plausible spatial audio rendering consistent with the real acoustic environment in order to create a coherent soundscape of real and virtual sources. In parallel, the rendering must be computationally lightweight, since the limited resources usually have to be shared with other computationally demanding components, such as visuals, sensors, and mapping, among others. One approach to meet these conflicting demands is parametric rendering. The sound field is first encoded offline into a parametric description covering the perceptually essential components and then decoded (in real-time) for efficient spatial audio reproduction scalable to the technical conditions. Furthermore parametric rendering offers a high degree of flexibility, as the parameters can be easily adjusted, for example, to represent different source-receiver conditions or room acoustic situations.

Various methods for the parametrization (encoding) and rendering (decoding) of sound fields based on spatial room impulse responses (SRIRs) have been presented (e.g., [1–6]). The methods usually split the sound field into directional and diffuse components, which are then processed and rendered separately. To determine the directional components, the approaches exploit the directional information of the SRIR to estimate the direction of arrival (DOA) of the direct sound and the early reflections.

The methods mentioned above usually decode the sound field only at the measurement point for the listener's head orientations, i.e., only for 3 degrees of freedom (DoF). Recent work extended those methods to render the sound field for 6 DoF and thus for arbitrary head orientations and room

*The work was done as a research intern at Facebook Reality Labs Research in Redmond, WA, USA.

positions of the listener. The methods typically require at least one or even multiple SRIRs from different positions in the room and derive a description of the sound field at arbitrary positions by extrapolation based on one SRIR (e.g., [7, 8]) or by interpolation between the distributed SRIR measurements (e.g., [9, 10]). For a comprehensive overview of 6-DoF rendering methods please refer to [10].

In the context of AR, however, it may be necessary to obtain a spatial parametric description of the sound field for 6-DoF rendering based on one monaural room impulse response (RIR). For example blind system identification used to estimate room acoustic parameters in real-time results in a parametric description of the monaural RIR, which needs to be further processed and finally decoded for parametric spatial audio reproduction. Advances in blind system identification could in the near future enable RIR estimation with consumer devices. Furthermore a monaural RIR can be measured relatively easily even with consumer equipment (such as a smartphone), which, if appropriately encoded and decoded, would allow users and content creators of AR applications to quickly and easily obtain a spatial audio reproduction corresponding to the real space. Moreover an enormous number of monaural RIRs are available to the public, which can be used to build a database of parametrically described spaces suitable for real-time AR audio rendering.

Encoding a monaural RIR provides common parameters such as the amplitude and the time of arrival (TOA) of direct sound and early reflections as well as the frequency-dependent reverberation time or the direct-to-reverberation ratio (DRR). However, since a monaural RIR does not contain directional information, the DOAs of the direct sound and the early reflections must, for example, either be predefined, pseudo-randomly assigned, or estimated based on the room geometry.

Pörschmann et al. [11] presented a first approach for synthesizing binaural room impulse responses (BRIRs) for any desired head orientation based on one measured RIR. Their method aims to decompose the broadband RIR into a directional and diffuse part and process them separately (quite similar to spatial impulse response rendering (SIRR), where several band-passed parts of a SRIR are decomposed into directional and diffuse components [1]). The directional part, consisting of direct sound and (grouped) early reflections, is estimated by reflection detection and synthesized by convolving small chunks of the RIR containing directional information with head-related impulse responses (HRIRs). The algorithm assigns a predefined DOA (and optionally TOA) to the direct sound according to the measurement setup while it assigns pseudo-randomized DOAs to the early reflections due to missing spatial information. The binaural diffuse reverberation is synthesized by convolving small chunks of the RIR with chunks of binaural white noise and summing with overlap-add, which is essentially a de-correlation of the RIR.

This paper presents a novel approach for parametric spatial audio with 6 DoF based on one monaural RIR. The method, which we named *Paraspax* (PARAmetrization, SPAtialization, and eXtrapolation of monaural room impulse responses), is inspired by the method introduced by Pörschmann et al. but has significant extensions and improvements.

Through encoding of the monaural RIR, the present method derives monaural and spatial parameters suitable for parametric BRIR synthesis or real-time parametric rendering. The approach provides a scalable reflection detection, which allows selecting a specific number of perceptually salient reflections [12] for processing and position-dynamic rendering. The DOAs for the selected early reflections can be derived in different ways: a) based on a pseudo-randomized directional distribution; b) based on a simple image source model (ISM) for a shoebox-shaped room, which approximates the real room's geometry; or c) based on a previously determined DOA pattern obtained, for example, by applying the spatial decomposition method (SDM) [3] to SRIR measurements. Thus the latter option makes it possible to combine SRIR measurements with the scalable encoding of the presented approach.

Furthermore the proposed method allows filtering the direct sound according to the sound source directivity, thus enhancing the presentation when, for example, walking around a virtual sound source. Assuming a shoebox-shaped room with the selected early reflections as image sources, the parameters can be extrapolated to any position in the room, which allows 6-DoF parametric spatial audio rendering.

We further present a processing chain for parametric BRIR synthesis based on the (extrapolated) parameters and measured monaural RIR. The method decomposes the sound field into directional and diffuse components. In contrast to previously described approaches, which also perform a decomposition (e.g., [1, 2]), the *Paraspax* method does not require spatial information (i.e., no multichannel RIR) for the decomposition of the sound field but achieves it solely based on a monaural RIR. In the synthesis, BRIRs are constructed by recombining the directional and diffuse components, which are adjusted according to the parameters.

In the paper we evaluate the synthesized BRIRs by comparison with measurements. Moreover we present an implementation of a 6-DoF virtual acoustic environment (VAE) in a selected room, based on BRIRs synthesized with the *Paraspax* method and a purpose-built real-time framework, which can be used to perform psychoacoustic experiments for research on AR. Lastly we present results of a pilot study using the 6-DoF framework to examine the perceptual plausibility of the *Paraspax* method.

1 *PARASPAX* METHOD

Paraspax can be grouped into three basic processing blocks for encoding: parametrization, spatialization, and extrapolation, as shown in Fig. 1. The parametrization provides standard monaural room acoustic parameters, the amplitude and TOA of direct sound and early reflections using a reflection detection algorithm, the magnitude responses of reflection filters, and the reverberation level. The spa-

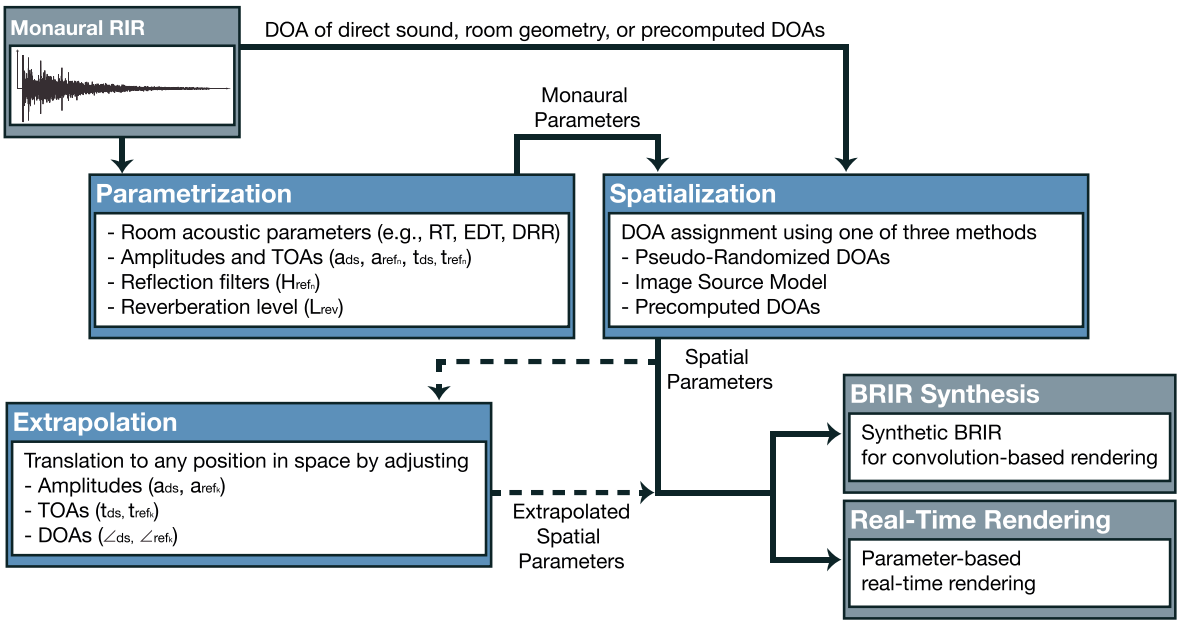


Fig. 1. Block diagram of the *Paraspax* method. The parametrization of the monaural RIR provides basic monaural room acoustic parameters, amplitudes and TOAs of the direct sound and early reflections, the magnitude responses of the reflection filters, and the reverberation level. The spatialization assigns DOAs to the direct sound and the selected reflections using one of the three implemented methods. The optional extrapolation allows for listener translations by adjusting amplitudes, TOAs, and DOAs of direct sound and early reflections. The parametric description can be used for BRIR synthesis or real-time parametric rendering.

tialization assigns DOAs to the direct sound and selected early reflections, based on a pseudo-randomized directional distribution, a simple ISM, or precomputed DOAs. Extrapolation allows listener translations to any position in space by adjusting amplitudes, DOAs, and TOAs of direct sound and early reflections. The parametric description can then either be used to synthesize BRIRs or passed to a real-time rendering engine. The following sections describe the encoding and BRIR synthesis in detail. A Matlab-based implementation of the proposed method with sample code is available online.[1]

1.1 Parametrization

Standard room acoustic parameters according to ISO 3382-2 are calculated in a first step based on the pressure response p, i.e., the monaural RIR, such as reverberation time ($RT_{20/30/60}$), early decay time (EDT), clarity ($C_{50/80}$), directness ($D_{50/80}$), and energy decay curves (EDC) in octave or 1/3-octave frequency bands, as well as the broadband DRR and the mixing time. The latter is calculated according to the approach proposed by Abel et al. [13], which follows the assumption that the sound pressure amplitudes in a diffuse sound field assume a Gaussian distribution. For this the normalized echo density profile is calculated (window length of about 21 ms, as recommended by the authors), which describes to what extent the amplitude distribution of an RIR approximates a Gaussian distribution over time. As diffuse energy increases, the echo density profile of an RIR rises, and the mixing time is defined as when the echo density profile reaches the value of one the first time.

[1] Available: https://github.com/facebookresearch/Paraspax

In further processing described in the following sections, the TOA and amplitude of direct sound (t_{ds}, a_{ds}) and the n detected early reflections (t_{ref_n}, a_{ref_n}), as well as the magnitude response of the reflection filters for each detected reflection (H_{ref_n}) and the reverberation level (L_{rev}), are estimated.

1.1.1 Direct Sound

The TOA of the direct sound t_{ds} is determined using onset-detection with a threshold of −20 dB in relation to the maximum value of the 10 times upsampled RIR. The amplitude of the direct sound a_{ds} is calculated as the root-mean-squared (RMS) average of an asymmetrical window centered around t_{ds}, starting 0.5 ms before and ending 1 ms after t_{ds}, as proposed by Brinkmann et al. [12]. Starting slightly before t_{ds} accounts for pre-ringing artifacts, and the length of 1 ms after t_{ds} relates to the time frame in which summing localization takes place, i.e., the time in which multiple coherent sound sources are perceived as one auditory event [14, ch. 3.1].

1.1.2 Early Reflections

The reflection detection can be performed on the entire RIR or in a limited time range, e.g., up to twice the calculated mixing time, to estimate the TOAs of the early reflections only in the early part of the RIR. The selection process is motivated by perceptual mechanisms and is described below.

The RIR is windowed with a sliding rectangular window of 1 ms, and the TOA of a reflection is defined as the time index where the local energy is 3 times higher than the median energy in the window [15]. The window size

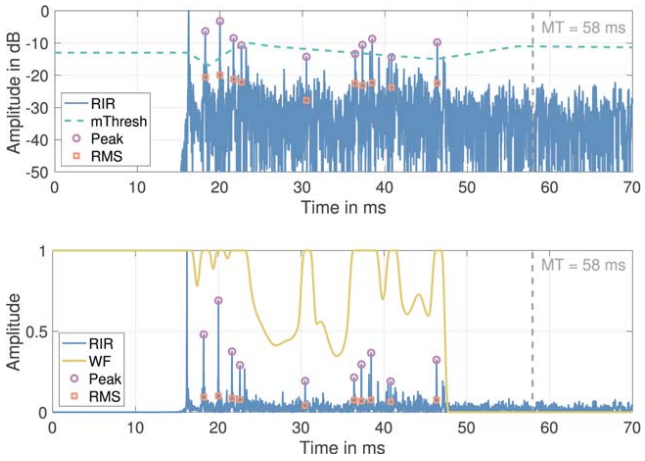


Fig. 2. Result of the reflection detection with $k = 10$ loudest reflections, peak and RMS amplitude values, adapted masking threshold (top), and estimated weighting function (bottom) for the RIR decomposition.

ensures a high temporal resolution in order to capture the perceptually important floor reflection [16, 17]. For each detected reflection, the RMS amplitude a_{ref_n} in an asymmetrical window of 1.5 ms (similar as described above for the direct sound), as well as the peak amplitude, is calculated.

The reflection selection method can lead to multiple detected reflections that are very close in time. To avoid this a next step evaluates whether another reflection occurs in a time range of 1 ms after a detected reflection (we choose this time range again according to mechanisms of summing localization). If this is the case, the reflection with the higher amplitude is declared valid, while the reflection with the lower amplitude is excluded from the selection. Optionally, in a further selection step, the peak amplitude of the detected reflections can be compared to the reflection masking threshold determined by Olive and Toole [18]. This procedure links the reflection detection even more closely to auditory perception and makes it possible to exclude potentially inaudible reflections from dynamic rendering.

Finally the selected reflections are sorted according to their amplitude and the k loudest early reflections are selected for spatialization and dynamic reproduction. Although other sorting and selection mechanisms (e.g., by time order) are also conceivable, previous studies have shown that rendering the loudest k (with $k = 6$–10) reflections is a valid method to reproduce the most perceptually important and salient reflections [19, 12].

Fig. 2 shows an example of the described reflection detection and selection with $k = 10$ loudest reflections. In addition to the peak and RMS amplitude values of each selected reflection the plot at the top shows the masking threshold [18] adapted to the RIR. To better estimate the audibility of reflections that arrive significantly later after the direct sound, the masking threshold could be further adjusted iteratively as a function of time [12]. However the current implementation seems to be a simple measure to validate whether the detected reflections (primarily the first reflections after the direct sound) are relevant. In the example shown in Fig. 2 it is interesting to see that all selected reflections are before the mixing time (MT = 58 ms). This indicates that at least in this example, the mixing time is a reasonably good predictor for the transition from a directional to diffuse sound field.

The plot at the bottom of Fig. 2 additionally shows the weighting function (WF) derived from the selected reflections, which is used in the BRIR synthesis (see SEC. 1.4) to decompose the RIR into a directional and diffuse part. To build the WF the envelope of the pressure response p is constructed by convolving its absolute value with a Hann window of 3 ms. This envelope function is set to 1 for the windows of 1.5 ms around the TOAs of the direct sound and selected reflections. Finally, the function constructed in this way is smoothed with a 1-ms window to avoid too-strong edges, which leads to the WF shown in Fig. 2 (bottom) as an example.

1.1.3 Reflection Filters

In addition to TOA, DOA, and amplitude, early reflections can be described more precisely by their spectral characteristics. Parametric rendering that, unlike the presented BRIR synthesis, does not use signal components of the RIR (e.g., [4, 19, 12]) requires so-called reflection filters to adjust the magnitude spectra of (synthetic) early reflections according to the frequency-dependent absorption properties of the reflecting surfaces. Gaining broadband reflection filters from a monaural RIR, however, can be problematic because windows of sufficient length (e.g., a window of 10 ms is required to analyze frequencies down to 100 Hz) around a reflection often contain other reflections that arrive later. The result is an inaccurate reflection filter with a comb-filter magnitude response consisting of parts of multiple reflections.

To still derive useful reflection filters, the proposed encoding determines the magnitude response of each selected reflection in three asymmetric windows of different sizes. The first window is equal to the 1.5-ms direct sound window as described in SEC. 1.1.1 and thus can analyze frequencies down to about 667 Hz. The longest third window has a length defined by the predefined lowest frequency to be resolved, e.g., 10 ms for a boundary frequency of 100 Hz. The length of the second window is then calculated to have a boundary frequency between the boundary frequencies of windows one and three, i.e., in the given example the second window would have a length of about 2.6 ms, leading to a boundary frequency of about 384 Hz.

The actual magnitude response is then composed of the different frequency components that can be resolved by the respective window. By this, most reflections are described correctly, at least in the higher frequency range (f > 667 Hz for the 1.5 ms window), whereas in lower frequency ranges, comb-filter structures can occur again, since these frequency components are based on longer windows.

To compensate for the spectral influence of the sound source, the magnitude responses of the reflections are filtered with the inverse magnitude response of the direct sound, which is also constructed from the frequency com-

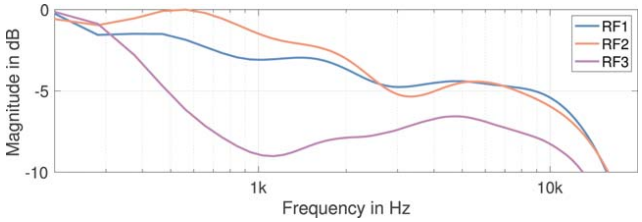


Fig. 3. Magnitude response H_{ref_n} of the estimated reflection filters for the first three loudest reflections.

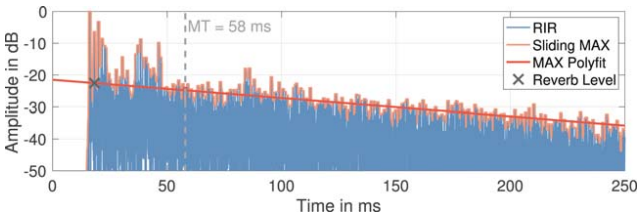


Fig. 4. Estimation of the reverberation level L_{Rev} (gray cross), defined as the amplitude value of the decay curve (red line) at the TOA of the first selected early reflection.

ponents of the same three different windows. The procedure of inverse filtering is quite similar to the estimation of absorption coefficients from in situ measurements [20] and provides reflection filters that can be used for parametric rendering independent of the sound source used for the RIR measurement.

The reflection filters determined this way are nevertheless only a rough approximation of the absorption properties of the reflecting surfaces and become more and more inaccurate as the RIR progresses in time and reflection density increases. To compensate for those inaccuracies to a certain degree, the reflection filters are octave-smoothed in a final step. Fig. 3 shows the magnitude response H_{ref_n} of three reflection filters as an example, obtained with the described method from the first three reflections ($n = \{1, 2, 3\}$) of the RIR shown in Fig. 2. Depending on the application, H_{ref_n} can be transformed by frequency sampling into linear or minimum-phase FIR filters, or for example, with the Yule-Walker method into IIR filter coefficients [21].

1.1.4 Reverberation Level

The last encoding step consists of the estimation of the reverberation level L_{Rev}. It describes the level of the diffuse sound field at the TOA of the first selected early reflection in the early directional part of the RIR. Reverberation level is an important parameter when synthesizing BRIRs using both a directional and diffuse reverberation component, since they must be combined at a certain level ratio (see, e.g., [4, 22]) to preserve the correct energy of the RIR.

In principle the DRR also describes such a level ratio, but since the energy is integrated over the entire length of the impulse response when calculating the DRR, the result depends on the length of the impulse response and thus also on the reverberation time. The proposed reverberation level parameter is independent of the reverberation time, so it is more convenient to use in practice, such as generating BRIRs with different reverberation times but similar level ratios between the directional and diffuse components. Furthermore the parameter can be used to estimate the diffuse energy in the early directional part or the amplitude at the mixing time for designing linear or cosine-squared ramps to fade in the diffuse reverberation (see, e.g., [19, 22]).

Fig. 4 shows an example of the calculation of the reverberation level. First the envelope of the absolute pressure response $|p|$ is estimated by calculating the maximum in a sliding window of 1 ms. The decay curve in decibels is then approximated by a first-order polynomial fit to the envelope in the time range between two and three times the mixing time. The selected time range ensures that there are no early reflections that distort the envelope, which would bias the determined decay curve. Lastly the decay curve is linearly extrapolated to obtain values for the time range before two and after three times the mixing time. The reverberation level is then defined as the amplitude value of the decay curve at the TOA of the first selected early reflection, which in this example is about -22.5 dB. Alternatively the amplitude of the reverberation at the mixing time can be determined by evaluating the decay curve at that time.

1.2 Spatialization

For spatialization, i.e., for assigning DOAs to the direct sound ($\angle_{ds}$) and the k selected early reflections ($\angle_{ref_k}$), the proposed method offers three different possibilities. The following sections explain the three approaches in detail.

1.2.1 Pseudo-Randomized DOAs

The first approach is inspired by the method introduced by Pörschmann et al. [11]. To obtain a correct representation of the direct sound its DOA must be known. The DOAs for the selected early reflections are based on a pseudo-randomized directional distribution stored in a lookup table. This lookup table with the DOAs can be derived, for example, from a shoebox-shaped room with non-symmetrically arranged source-receiver positions.

It is evident that the DOAs assigned this way match the actual reflection pattern only to varying degrees, depending on the situation. However, listening experiments showed that synthetic BRIRs with pseudo-randomized DOAs show quite high perceptual similarity with measured reference BRIRs [11, 23]. Thus, depending on the application, and if no further information about the sound field or space is available, such spatialization may be sufficient. Nevertheless the listening experiments by Pörschmann et al. [11] also showed that spatial attributes, such as apparent source width, localizability of the sound source, or listener envelopment, are significantly impaired by assigning incorrect DOAs. Thus more accurate spatialization methods should be preferred when possible.

1.2.2 Image Source Model

If the approximate dimensions of the room (length, width, height) and the source and receiver positions are

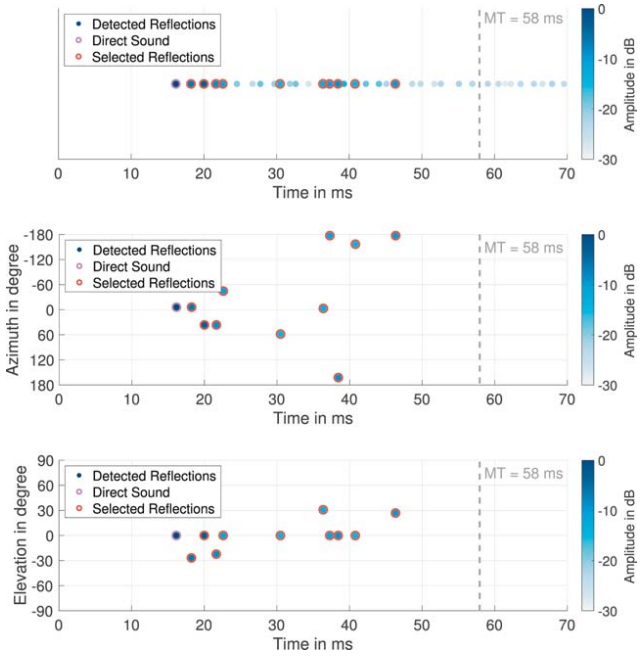


Fig. 5. Result of the spatialization based on a second-order image source simulation. Amplitude and TOA of the direct sound and the early reflections, acquired from the monaural RIR with the presented reflection detection (top). Spatialized direct sound and the selected $k = 10$ loudest reflections with assigned DOAs described by azimuth (middle) and elevation (bottom).

known, more appropriate DOAs for the selected early reflections can be obtained using an image source simulation for a shoebox-shaped room, which approximates the actual geometry of the room. The order of the image source simulation is freely selectable but since perceptual evaluations of parametric rendering have shown that six dynamically reproduced early reflections (theoretically the first-order reflections in a shoebox room) are already sufficient [19, 12], a second-order simulation is usually enough. In this spatialization mode, the DOA and distance of the direct sound can either be predefined or obtained from the simulation.

To estimate the DOAs of the selected early reflections, the TOAs estimated with the reflection detection (see SEC. 1.1.2) are compared with the TOAs calculated with the simulation, and reflections with equal TOA or smallest TOA differences are paired. The selected early reflections are then assigned the DOAs of the corresponding simulated reflections. The algorithm can be set first to assign the more important first-order reflections and then, in a second step, the second-order reflections to the remaining estimates. The described approach allows a relatively simple but much more correct spatialization than using pseudo-randomized DOAs, provided that the room's geometric information is available or can be determined.

Fig. 5 shows an example of the spatialization based on a second-order image source simulation. The plot at the top shows the TOA and amplitude of the direct sound and all detected early reflections, as extracted with the proposed encoding (see SECS. 1.1.1 and 1.1.2) from the RIR presented in Fig. 2. The two other plots show the azimuth (middle plot) and elevation (bottom plot) of the spatialized direct sound and the $k = 10$ loudest early reflections with assigned DOAs that are dynamically updated during rendering.

1.2.3 Precomputed DOAs

A precomputed DOA pattern can also be used for spatialization, making use of multichannel RIRs to derive DOA estimates. For example DOAs estimated in the context of SDM with open microphone arrays exploiting time differences of arrival (TDOAs) or with B-format array using pseudo intensity vectors (PIVs) [3, 24, 1, 25] can be passed to the spatialization. The direct sound and selected reflections are then assigned the DOAs listed according to the TOAs in the passed DOA vector. In the best case the DOA vector is post-processed, i.e., smoothed and with stabilized direct sound [26], to improve the perceptual quality of binaural renderings.

Using precomputed DOAs is an efficient way of combining an accurate spatialization with the scalable encoding and decoding of the presented method. The parametrization enables efficient storage and rendering of a multichannel RIR and a handful of parameters. Additionally, with only one source-receiver combination, the results can be extrapolated to other positions.

1.3 Extrapolation

The monaural and spatial parameters are sufficient for 3-DoF rendering at the receiver position, i.e., dynamic spatial audio reproduction for any head orientation of the listener (yaw, pitch, roll), for example using dynamic binaural synthesis. In this case only the DOAs of the direct sound and the early reflections ($\angle_{\mathrm{ds}}$ and $\angle_{\mathrm{ref}}$) have to be dynamically adjusted according to the head orientation by applying a corresponding rotation matrix. However for 6-DoF rendering, i.e., when the listener moves through the room, the amplitude and the TOA of the direct sound and the early reflections (a_{ds}, a_{ref}, t_{ds}, and t_{ref}) must also be adjusted accordingly. Since there are no further measurement points in the present case we refer to this as an extrapolation of the parameters to the new listener position.

The extrapolation adjusts the amplitudes, TOAs, and DOAs according to an image source model, as shown in Fig. 6, exemplified by direct sound and two reflections. For this purpose the TOAs are transformed into distance values describing the distance to the sound source (direct sound) or the so constructed image sources (reflections), and in combination with the DOAs, they are converted into Cartesian coordinates. The extrapolated DOAs and TOAs are then obtained by subtracting the displacement vector, describing the listener translation, and transforming the result back to spherical coordinates. The amplitudes are adjusted according to the change in distance to the sound or image source using the inverse-square law.

The extrapolated set of parameters describes the directional sound field for frontal head orientation and can again be adjusted according to the listener's head orientation by rotating the DOA vector. As the level of the diffuse reverberation is kept constant in rendering or BRIR synthesis,

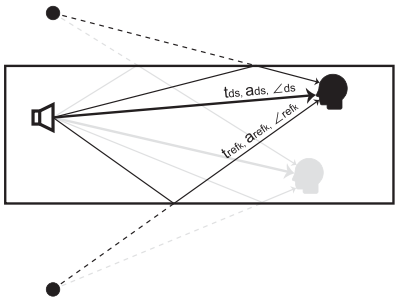


Fig. 6. Extrapolation to a new listener position (gray to black) by adjusting the amplitudes, TOAs, and DOAs of the direct sound and the (here shown as example two) early reflections based on an image source model.

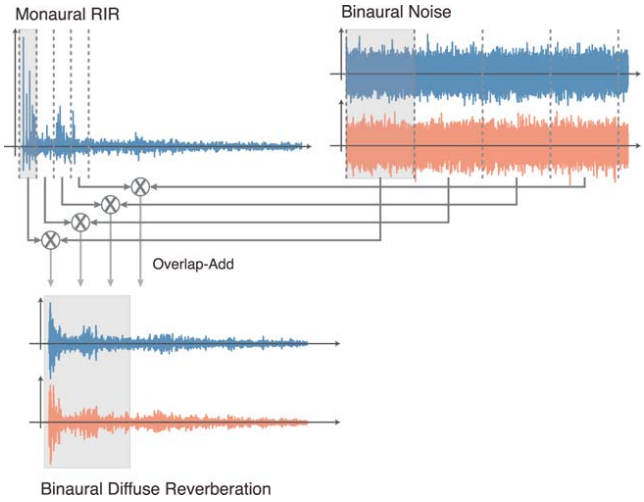


Fig. 7. Synthesis of the binaural diffuse reverberation by convolution diffusion. Small chunks of the RIR are convolved with small chunks of binaural noise and summed with overlap-add (adapted from [11]).

adjusting the amplitudes of the directional components and filtering the direct sound according to the source directivity also changes the DRR appropriately as the listener moves through the room (see SECS. 1.4 and 2). Thus two important cues for distance perception in enclosed spaces, level and DRR [27], are correctly reproduced.

1.4 BRIR Synthesis

The parametrization, spatialization, and optional extrapolation provide sufficient information for parametric BRIR synthesis or real-time parametric rendering. Thus completely synthetic BRIRs (i.e., BRIRs without any signal components of the original RIR) could be generated based only on the parameters using one of the many synthesis algorithms (see, e.g., [28, 29, 4, 22, 12]). However the synthesis that we have implemented as part of the *Paraspax* framework uses signal components of the monaural RIR in addition to the parameters to generate the BRIRs, with the great advantage that the essential sound characteristics of the room are preserved.

The diffuse components (the binaural diffuse reverberation) and directional components (the spatialized direct sound and early reflections) are first synthesized separately and finally added to form a BRIR. Besides the monaural RIR and parametric description (as parts of the monaural RIR are used for the synthesis, only the amplitudes, TOAs, and DOAs determined in encoding are required), the synthesis needs a full-spherical set of Head-Related Transfer Functions (HRTFs, i.e., the frequency-domain equivalent of the HRIR) and, if the directivity of the sound source is to be considered, corresponding directivity measurements. As part of a 6-DoF spatial audio system the synthesis can be used to precompute BRIRs for arbitrary head orientations and room positions, which can then be used for real-time convolution (see SEC. 3).

1.4.1 Diffuse Components

The binaural diffuse reverberation is generated based on the monaural RIR by convolution diffusion [1], i.e., the monaural RIR is filtered with binaural noise, which is quite similar to filtering with rectangular noise bursts in order to decorrelate omnidirectional reverberation for loudspeaker representation [30]. Fig. 7 illustrates the method implemented in *Paraspax*, which is essentially based on the convolution diffusion approach proposed by Pörschmann et al. [11]. First binaural noise is generated, which is two-channel white noise, filtered with the diffuse field response (also called common transfer function [31, ch. 1]) of the applied HRTF set and a coherence filter modeling diffuse-field interaural coherence (IC) [32]. The binaural diffuse reverberation is then synthesized by convolving small chunks of the RIR (0.67 ms; 32 samples at 48 kHz sampling rate) with chunks of the binaural noise (2.67 ms; 128 samples at 48 kHz sampling rate) and summing with overlap-add. This processing results in binaural reverberation with the frequency-dependent decay of the monaural RIR as well as with the diffuse-field IC and the associated spaciousness.

We determined the time constants for the RIR and the noise chunks with informal listening comparing to measured references. The listening revealed that longer noise blocks lead to a more spatial and less neutral-sounding reverberation, as also informally determined in a previous study [1]. Our informal evaluation showed that a block length of 32 samples for the RIR and 128 or 256 samples for the noise (at a sampling rate of 48 kHz) provides the best results.

1.4.2 Room Impulse Response Decomposition

Directional components mainly characterize the early part of a BRIR, where most energy is contributed from specular reflections. As the BRIR progresses in time, the relative contribution of diffuse energy progressively increases [33, ch. 4]. Listening experiments showed that mixing directional and diffuse components in the early part of a BRIR provides higher perceptual quality than BRIRs with purely directional components, indicating that synthetic BRIRs should exhibit diffuse sound also in the early part [4, 19].

The approach presented here synthesizes the early part of a BRIR as the superposition of weighted specular components and diffuse components. Fig. 8 (top) illustrates

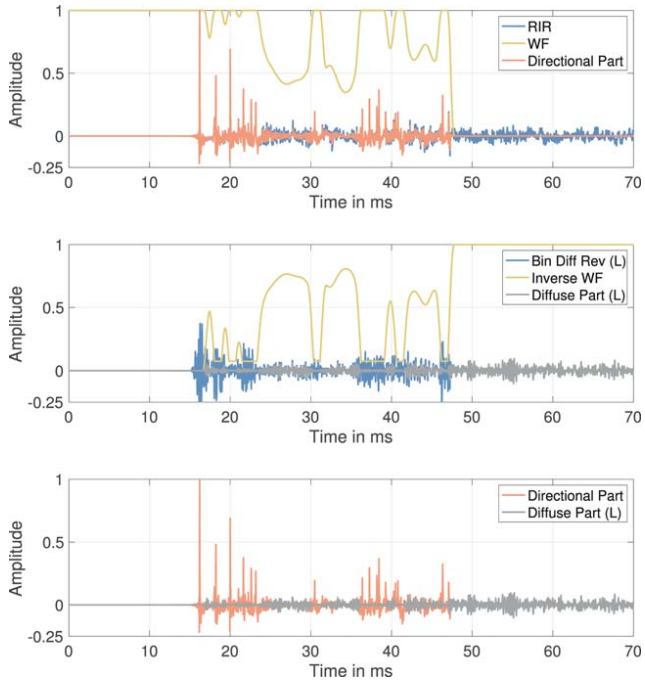


Fig. 8. RIR decomposition in the early part. Extraction of the directional (specular) components by applying the weighting function (top). Estimation of the diffuse components by applying the inverse weighting function to the synthesized binaural diffuse reverberation (middle). For illustration, superimposed monaural specular and (left channel) binaural diffuse part (bottom).

how the monaural RIR is decomposed to extract the directional (specular) components applying the weighting function (WF) estimated with the reflection detection (see SEC. 1.1.2). The extracted directional part is the basis for the synthesis of early reflections in the synthetic BRIRs (see SEC. 1.4.3).

Fig. 8 (middle) shows how the diffuse component in the early part is estimated by applying the inverse WF to the binaural reverberation—only the left channel is shown in the example. The inverse WF is constructed as the square root of the inverted weighting function, limited to the reverberation level estimated in encoding so that diffuse energy is also preserved in specular events (i.e., where WF is equal to 1). To illustrate the decomposition, Fig. 8 (bottom) finally shows the extracted monaural specular part overlayed with the left channel of the binaural diffuse part.

1.4.3 Directional Components

The direct sound and selected early reflections are the only components adjusted according to the listener's head orientation or position in the *Paraspax* synthesis, which allows generating BRIRs for 6-DoF spatial audio reproduction. The synthesis can be performed for any number of head orientations as well as for any listener position. The head orientations are represented by a spatial sampling grid with azimuth and elevation describing either the relative orientation to the sound source (global coordinate system) or the listener-related head orientation (local coordinate system), whereas the listener position is described by Cartesian coordinates in a global coordinate system with the origin being the measurement position of the monaural RIR.

The *Paraspax* algorithm performs the synthesis of the directional components in the following steps. First, chunks of the RIR are extracted in the non-symmetric 1.5-ms windows around the TOAs of the direct sound and the selected early reflections. When an extrapolated position is synthesized the chunks are adjusted in amplitude with the amplitude factors derived from the previous parameter extrapolation (see SEC. 1.3). If sound source directivity measurements are available a directivity filter is applied to the direct sound in the next step.

Since the sound source directivity is usually already imprinted in the direct sound extracted from the monaural RIR, the direct sound is filtered only according to the change of the directivity corresponding to a change of position relative to the sound source. Thus, for the synthesis at the measurement position, the directivity filter has a flat magnitude response over the entire frequency range (i.e., no filtering). However for extrapolated positions, the direct sound is filtered according to the direction and frequency-dependent change in sound source directivity. In the present case the directivity measurements are stored as spherical harmonics (SH) coefficients at a sufficiently high spatial order $N \geq 35$. Using the SH description allows artifact-free SH interpolation of the directivity to obtain suitable directivity filters for any radiation direction [34].

Next the (processed) RIR chunks are convolved with HRTFs according to the DOAs for the respective head orientation, resulting in a specific directional pattern for each point of the spatial sampling grid. To calculate the DOAs first estimated for frontal head orientation at the measurement or extrapolated position, a rotation matrix according to the head orientation is applied (see, e.g., [26]). As with the directivity measurements, the HRTF set is stored as SH coefficients at $N \geq 35$, allowing HRTFs to be obtained for any direction.

Next the resulting directional components (i.e., the convolution result of RIR chunks and HRTFs) are placed at their respective position in time according to the estimated or extrapolated TOAs. In a last step, the amplitude of the synthetic direct sound for frontal head orientation at the measurement position (azimuth $\phi = 0^\circ$ and elevation $\theta = 0^\circ$ relative to the source) is compared with the amplitude of the direct sound from the monaural RIR, and the level of the entire synthetic directional component is adjusted accordingly. This level adjustment, which is a constant factor applied to all synthesized directional components, compensates for level changes that may occur, for example, due to convolution with non-level normalized HRTFs, which would drastically change the DRR in the final BRIRs.

Fig. 9 shows as an example the result of the described synthesis of the directional components of a synthetic BRIR for the measurement position and frontal head orientation. The synthesis is based on the extracted direct sound and extracted ten loudest reflections of the monaural RIR shown in Figs. 2 and 8.

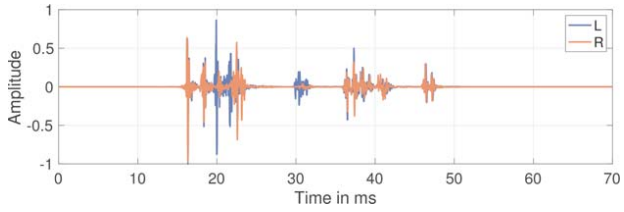


Fig. 9. Synthesized directional components.

1.4.4 Composition of Directional and Diffuse Components

To generate the final synthetic BRIRs, the directional components, which are two-channel BRIRs for each head orientation at each listener position, are combined with the diffuse components for the early part and with the binaural diffuse reverberation characterizing the late part of the BRIR. Fig. 10 (top left) shows the early part of the synthetic BRIR, which is a composition of the weighted synthetic binaural reverberation (see SEC. 1.4.2) and the directional components synthesized based on the specular part of the monaural RIR (see SEC. 1.4.3). The top right plot in Fig. 10 shows a longer segment of the same synthetic BRIR with a logarithmic amplitude scale to better illustrate the decay of the binaural reverberation.

As an example, the two plots at the bottom left and right in Fig. 10 show a synthetic BRIR for an extrapolated position (two meters back and one meter to the right from the measurement position) with the same head orientation. The amplitude and TOA shifts are clearly visible in the early part, whereas the late part is the same for all BRIRs.

The *Paraspax* toolbox offers various post-processing options to further filter the synthetic BRIRs. Hence filters can be applied to compensate at least to some extent for room modes that appear in the monaural RIR, to change the IC of the directional or diffuse components, or to high-pass the BRIRs or only the diffuse components to control low-frequency ($f \leq 50$ Hz) spaciousness and reverberation.

2 PHYSICAL EVALUATION

For the physical evaluation of the *Paraspax* synthesis we compared measured RIRs and BRIRs with synthesized BRIRs concerning various room acoustic monaural and binaural parameters. Even if the synthesis method does not attempt to physically reconstruct the room response, synthetic and measured impulse responses should ideally match within the limits of the respective just-noticeable differences (JNDs) of room acoustic parameters to achieve satisfactory perceptual results.

2.1 Room and Measurement Setup

We conducted measurements and implemented a 6-DoF spatial audio system (see SEC. 3) based on BRIRs synthesized with the *Paraspax* toolbox in a mostly empty shoebox room. Fig. 11 (top) shows a cross-section of the room with dimensions 11.73 m $\times$ 4.74 m $\times$ 4.62 m (length $\times$ width $\times$ height). We measured the impulse responses using the swept-sine technique with an Earthworks M30 as well as with a KEMAR dummy head on a 4 m $\times$ 3 m rectangular grid of 1-m resolution and a measurement height of 1.40 m. It is important to note that we measured the KEMAR BRIRs with Brüel & Kjær 4101 binaural in-ear microphones (placed in the ear canal of the KEMAR) and not with the microphones integrated in the dummy head. The sound sources were four Genelec 8020 loudspeakers at different positions in the room and different heights.

Fig. 11 (bottom) shows a picture of the room with the KEMAR dummy head (wearing AKG K1000 headphones) at position 10 of the grid, the 4 Genelec loudspeakers, and an optical tracking system (OptiTrack) covering the entire room. The room has plain walls, a large glass window, and a concrete floor, leading to distinct early reflections and a relatively high reverberation time of $RT_{30} = 0.9$ s (average between 0.25 and 8 kHz).

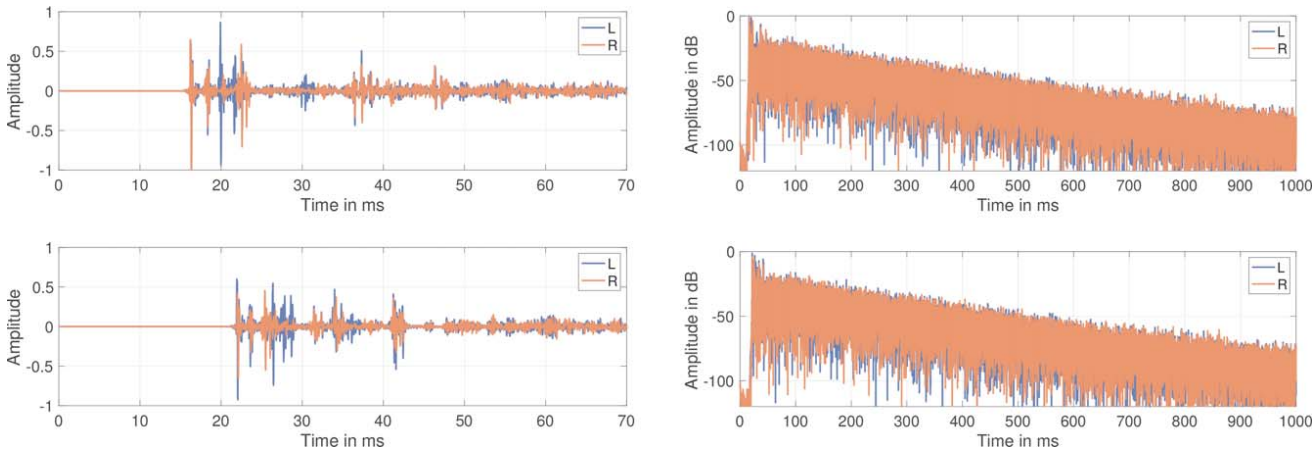


Fig. 10. Early part (left) and longer segment (right) of synthesized BRIRs, composed of directional components and binaural diffuse reverberation. The plots at the top show a synthesized BRIR for the measurement position of the monaural RIR; the plots at the bottom a BRIR for an extrapolated position (two meters back and one meter to the right from the measurement position).

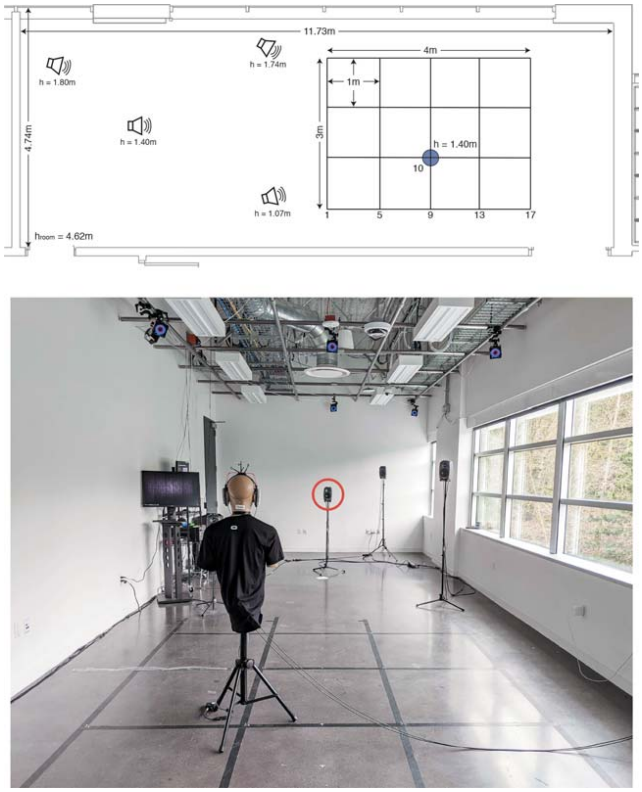


Fig. 11. Cross-section of the room with the positions of the four loudspeakers and the measurement grid/listening area (top). Picture of the room with a KEMAR dummy head at position 10 of the grid, the four Genelec 8020 loudspeakers, and the OptiTrack system (bottom).

2.2 Synthesized BRIRs

We have chosen position 10 as the reference position for this study, which means that all synthesized BRIRs for each position, head orientation, and loudspeaker for this room are based on 1 monaural RIR for each loudspeaker measured at position 10. Due to the high number of measurement points we limited the physical evaluation presented here to BRIRs synthesized for loudspeaker 2 (highlighted loudspeaker in Fig. 11 (bottom)), which has a height of 1.40 m and distance to position 10 of about 5.5 m. For comparison we synthesized BRIRs for frontal head orientation (local head-related coordinate system) based on the monaural RIR measured at position 10 with loudspeaker 2 as the source, with the spatialization based on a second-order image source simulation (see SEC. 1.2). The loudspeaker was slightly offset to the right in azimuth from position 10 (approximately 6°).

For the synthesis we used KEMAR HRTFs, measured with Knowles FG-23329 miniature microphones at the blocked ear canal of the dummy head. The measurements were done on a high-resolution spherical sampling grid (9,720 directions) in the anechoic chamber at Facebook Reality Labs Research. The HRTFs have been low-frequency corrected below 200 Hz [35] and transformed to SH domain at $N = 35$ using the discrete spherical Fourier transform with Tikhonov regularization [34]. The directivity data for the Genelec 8020 loudspeaker were taken from the BRAS database [36] and also transformed to the SH domain at $N = 35$.

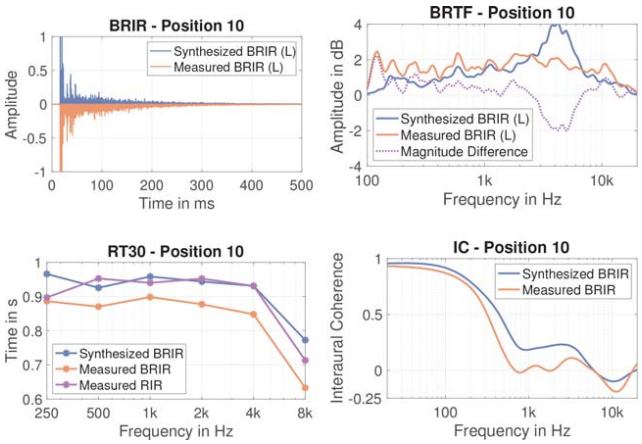


Fig. 12. Comparison of measured and synthesized impulse responses at position 10 in terms of time-energy structure (top left), magnitude response (top right), reverberation time RT_{30} (bottom left), and frequency-dependent IC (bottom right).

2.2.1 Measurement Position

Fig. 12 provides a first comparison of the measured and synthetic impulse responses at position 10. The synthesized and measured broadband pressure BRIRs for the left ear (top-left plot) show good similarity in their overall time-energy structure with matching amplitude and time events. However the measured BRIR shows a finer resolved early part, mainly because the synthesis is based only on the ten loudest reflections.

The left-ear binaural room transfer functions (BRTFs) in the top-right plot show a good agreement between 0.2 and 2 kHz with magnitude differences of 1 dB or less. However in the frequency range between 2 and 5 kHz, there are noticeable deviations in magnitude of about 2 dB. These variations may occur because the BRIRs were measured in a different way than the HRTFs used for synthesis, i.e., the BRIRs were measured with microphones in the ear canals (not fully blocked) and the HRTFs were measured with different microphones on the blocked ear canal. Furthermore the diffuse field response obtained from the employed HRTF set and applied to the synthesized binaural reverberation (see SEC. 1.4.1) may increase slightly too much between 3–6 kHz.

The plot at the bottom left of Fig. 12 compares the reverberation times of the different impulse responses, where RT_{30} of the BRIRs was calculated as the mean value of RT_{30} for the left and right ear. The plot reveals that the synthesized BRIR and measured RIR have nearly identical frequency-dependent reverberation times, indicating that the proposed RIR decomposition and reconstruction when synthesizing BRIRs works well. Furthermore it shows that the presented convolution diffusion approach is well suited for synthesizing binaural reverberation from a monaural RIR while maintaining the decay of it.

The reverberation time estimated from the measured BRIR is slightly lower, possibly due to the dummy head acting as a directional receiver. To further investigate this we estimated the reverberation time for BRIRs measured

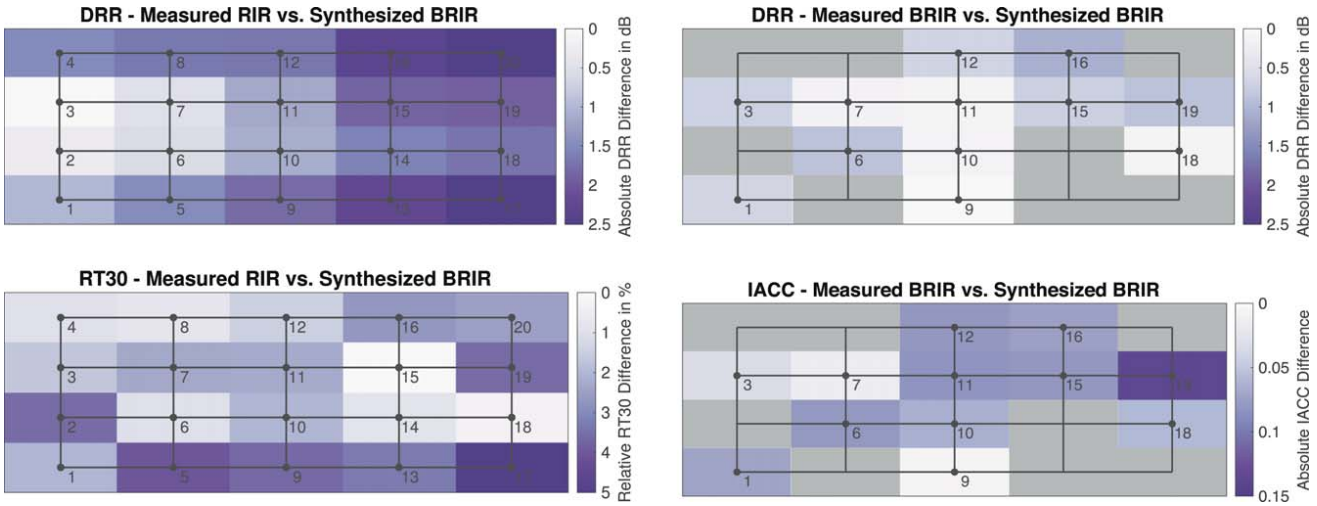


Fig. 13. Differences in DRR (top left and right), RT_{30} (bottom left), and IACC (bottom right) between measured and synthesized impulse responses at the grid positions. The colored rectangular fields visualize the magnitude of the differences at the grid positions, which are represented by a small dot in the center of the respective rectangular field. The gray fields (top and bottom right) denote positions without measurement data.

with different head orientations, confirming that the reverberation time varies slightly depending on the head orientation. Overall the reverberation times estimated from the measured BRIRs (for frontal head orientation) are lower at all positions in the room (see Table 1 in the APPENDIX). However as the synthetic BRIR is based on the measured RIR it is sensible to expect that in the present evaluation the reverberation times of these two impulse responses are mostly identical. Further studies are required to examine the effect of head orientation on estimated reverberation time, as well as the audibility of these deviations.

Lastly the bottom-right plot shows the frequency-dependent IC of the measured and synthesized BRIRs [28]. The plot reveals a very similar IC over frequency, although the synthesized BRIR has a slightly higher IC than the measured BRIR, especially above 200 Hz. These deviations indicate that the measured BRIR has more incoherent directional events than the synthesized BRIR.

2.2.2 Extrapolated Positions

To further examine the performance of the presented 6-DoF synthesis we compared DRR, RT_{30}, and Interaural Cross-Correlation (IACC) of synthesized and measured impulse responses at different points on the measurement grid. As mentioned above, all synthetic BRIRs were based on the monaural RIR measured at position 10 and were generated for frontal head orientation (local head-related coordinate system) by applying the extrapolation and synthesis described in SEC. 1. Accordingly the KEMAR BRIRs were also measured for frontal head orientation at the different positions in the room (see also Fig. 11).

The DRR was calculated with a direct sound window of 1.5 ms and for two-channel BRIRs, the direct and diffuse sound energy of the left and right channels were summed before calculating the DRR [37]. The broadband RT_{30} was calculated as the mean value of the reverberation times in the octave bands from 0.25–8 kHz. In the case of BRIRs the reverberation times determined per channel were averaged to obtain a single value.

Table 1 in the APPENDIX lists the individual DRR, RT_{30}, and IACC values for the respective impulse responses and grid positions. For better comparability off the estimated values, the plots in Fig. 13 show the deviations in DRR, RT_{30}, and IACC between measured and synthesized impulse responses at the grid positions. The colored rectangular fields visualize the magnitude of the differences at the measurement positions, with each measurement position represented by a small dot in the center of the respective rectangular field.

The absolute DRR differences between measured RIRs and synthesized BRIRs (top-left plot) exceed 2 dB only at a few positions (points 13, 16, 17, and 20), with a maximum of about 2.6 dB at position 20. Larsen et al. [38] determined JNDs of about 2–3 dB in rooms with a DRR of 0 or 10 dB and JNDs of about 6–8 dB in rooms with a DRR of −10 or 20 dB. In the present case the DRR ranged between −2.5 to −11 dB for the monaural RIRs, and thus the absolute DRR differences are well below estimated JNDs for DRR changes.

However as there are generally significant differences between the DRR estimated from a BRIR or from an omnidirectional RIR, that is, due to the directivity of the artificial head or due to the influence of the source angle [37], we have also compared the DRR of synthesized and measured BRIRs at different positions in the room. The results presented in Fig. 13 (top right) reveal absolute DRR differences far below JND, with a maximum difference of only about 1 dB at position 16 and differences of even less than 1 dB at all other evaluated positions. Please note that we have only made KEMAR measurements at those positions where numbers appear in the plot and have grayed out the rectangular fields without data.

The bottom-left plot in Fig. 13 shows the relative RT_{30} difference between measured RIRs and synthesized BRIRs

at all 20 positions. The largest deviation occurs at position 17, with about 5% relative difference between the broadband reverberation times of the synthesized BRIR and monaural reference RIR. At all other positions the relative difference is even below 5% and thus clearly below the JND for reverberation, which ISO 3382-1 lists as 5% according to Seraphim [39]. When considering the RT_{30} differences, however, it should be noted that all BRIRs use the same binaural reverberation synthesized from the monaural RIR measured at position 10, and therefore the reverberation time of the BRIRs is nearly identical at all points in the room. Thus the analysis presented here rather gives information about how homogeneous the sound field is in the room and how well the criterion of diffusity, which we indirectly apply in the synthesis, is fulfilled.

Lastly, Fig. 13 (bottom right) shows the absolute IACC differences between synthesized and measured BRIRs at 12 positions in the room. The differences exceed the JND for IACC, which is defined as 0.075 according to ISO 3382-1, only at position 19. At positions 6, 11, 12, and 15 the differences are in the JND range, whereas at all other positions the differences are clearly below the JND. Overall the broadband IACC for the KEMAR BRIRs ranges from 0.16 to 0.40 and from 0.14 to 0.43 for the synthesized BRIRs.

Interestingly the broadband IACC of the synthesized BRIRs is often slightly higher than that of the measured BRIRs, which is in line with the observations for the frequency-dependent IC at position 10 (see Fig. 12). These findings indicate that the measured BRIRs contain more directional and incoherent components (i.e., early lateral reflections). However as the differences exceed the JND for IACC only at one position we assume that the perceived spaciousness is similar throughout the room, regardless of whether synthetic or measured BRIRs are used for 6-DoF spatial audio reproduction.

3 SIX-DEGREES-OF-FREEDOM VIRTUAL ACOUSTIC ENVIRONMENT

The physical evaluation based on room acoustic parameters showed a good performance of the *Paraspax* method. However the study's main goal was implementing a perceptually plausible 6-DoF VAE based on one monaural RIR, which can be used, for example, for perceptual studies on AR audio. For this reason we implemented a 6-DoF framework using precomputed BRIRs synthesized with the *Paraspax* toolbox, with which demo applications and plausibility studies according to Lindau and Weinzierl [40] can be performed.

3.1 Real-Time Framework

The block diagram in Fig. 14 gives an overview of the 6-DoF framework, deployed in the room where we also performed the physical evaluation (see Fig. 11). A Matlab software developed for the framework takes over the central control. The software receives tracking data from an optical

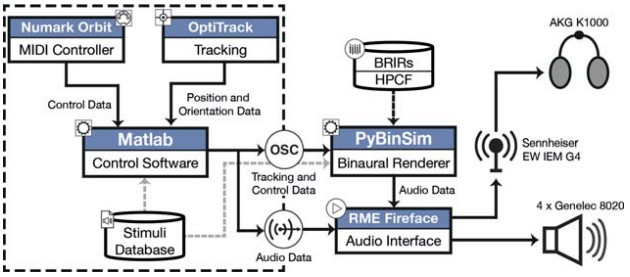


Fig. 14. Block diagram of the implemented 6 DoF real-time framework for demo applications and psychoacoustic experiments in AR audio.

tracking system (OptiTrack with 7 Prime 41 cameras) installed in the room at an update rate of 150 Hz and control data from a Numark Orbit MIDI controller at an update rate of 30 Hz. The OptiTrack system provides the position of the four Genelec 8020 loudspeakers and the position and orientation of the listener, and the Matlab software calculates from these data the relative orientation of the listener to the loudspeakers (azimuth, elevation, and distance) and the listener's absolute position in the room (X, Y, Z).

In the demo application, the listener can use the wireless MIDI controller to switch between real and virtual sources, switch between 1 of the 4 (virtual) sources, switch randomly between 1 of the 50 loudness-normalized test stimuli, or stop or pause playback. These control options enable a detailed comparison between real and virtual sound sources while moving through the room. In the case of a plausibility study the participant uses the MIDI controller to enter the answers. The Matlab software sends tracking and control data via OSC to the binaural renderer and audio data to a RME Fireface UCX audio interface if a real source (i.e., a loudspeaker) is played.

As the binaural renderer we used PyBinSim, a Python tool for real-time dynamic binaural synthesis [41]. The renderer convolves the dry audio signal with BRIRs, which have to be precomputed according to the desired spatial resolution for each head orientation and position in the room. The renderer allows using BRIRs split into an early and late part. In this case, only the early (directional) part needs to be adjusted according to the listener's head orientation and position by switching the early BRIRs, whereas the late part, i.e., the binaural diffuse reverberation, which is the same in all BRIRs, is covered by a single reverberation BRIR. This approach reduces the required memory and improves performance.

The BRIRs for each of the 4 sources were all based on 1 monaural measurement at position 10. For encoding of the four monaural RIRs, we used the *Paraspax* method as described in SEC. 1, with the spatialization based on a second-order image source simulation (see SEC. 1.2). We synthesized BRIRs using the method described in SEC. 1.4 with a spatial resolution of 4° in azimuth and 10° in elevation (restricted to ± 50°), resulting in 990 BRIRs per position. The uniform grid resolution, that is, the extrapolation steps in X and Y direction, was 0.25 m, which leads to 285 points on the 4 m × 3 m listening area. Given the

four different sources, a total of 1,128,600 (early) BRIRs were rendered.

To enable a direct comparison between real and virtual sound sources we used the extraaural AKG K1000 headphones for binaural playback, which were equalized with minimum phase FIR headphone filters designed with automatic regularization [42]. For better mobility the headphone signal was transmitted via a Sennheiser EW IEM G4 wireless in-ear monitor system. Furthermore the real sources (i.e., the loudspeakers) and their virtual representation over headphones were matched in loudness so that no level changes occurred when switching between real and virtual sources. The overall playback level in the listening area was $L_{\mathrm{Aeq}} = 60$–70 dB, measured for the closest sound source (i.e., loudspeaker 1).

The latency of the system was estimated to be approximately 66 ms, which is within the range of empirically determined thresholds of just detectable latency of 60–75 ms [43], and thus should be low enough in most real-life cases. The system latency consists of approximately 41 ms latency by the OptiTrack system [44], approximately 7 ms latency by processing the tracking data in Matlab (update rate of 150 Hz), approximately 6 ms due to OSC latency [45], about 11 ms due to the 512 samples buffer size in PyBinSim ($f_s = 48$ kHz), and additionally about 1 ms due to the RME Fireface UCX running with 64 samples buffer size ($f_s = 48$ kHz).

For a similar system (same tracking and playback system but a different binaural renderer), a similar motion to sound latency of about 70 ms was measured [46], suggesting that the latency estimate for the present 6-DoF VAE is quite accurate. However some participants in the plausibility experiment (see SEC. 3.2) could perceive head-tracking latency during high-speed rotations, suggesting that the actual latency of the system was higher than estimated. Accordingly latency measurements and possibly optimization of the system to reduce latency is inevitable in future work.

3.2 Plausibility Experiment

To assess the plausibility of the presented 6 DoF parametric spatial audio system we implemented a two-alternative forced-choice (2AFC) experiment as proposed by Lindau and Weinzierl [40]. During each trial a real or virtual source was randomly presented to the participants and the participant had to answer whether it was a real sound source (i.e., the sound came from one of the four loudspeakers) or a virtual sound source (i.e., the sound was reproduced over headphones).

The answers were analyzed with signal detection theory (SDT), i.e., the percentages of correct answers resulting from the 2AFC test were transformed to the criterion-free measure for the sensory difference d' between the presentation of a real or virtual sound source. A d' of 0 (50% correct answers; guessing rate) indicates inaudible differences, i.e., the VAE is perfectly plausible, whereas $d' > 0$ suggests audible differences between the presentations. A common critical value above which an auralization is considered to be no longer plausible is $d' = 0.82$, which corresponds to a 2AFC detection rate of $p_{2\mathrm{AFC}} = 0.72$ [47].

Each run had 120 trials consisting of 30 different test signals and 4 sound source positions. Thus, each of the 30 test signals was played exactly once with each of the 4 sources, whereby half of the trials were real and half were virtual sound sources. The order of test signal, sound source, and real or virtual presentation was randomized for each participant. The test signals were loudness normalized monophonic audio content with a length of 10 s. It included white and pink noise bursts, male and female speech, vocals, wind, string, percussion instruments, and synthesizer sequences.

The high number of different test signals prevents familiarization and the fact that the same combination of a test signal and sound source position is never presented both as a real and virtual source corresponds to the basic idea of plausibility testing [40, 47]. This procedure clearly distinguishes the plausibility test from a test on authenticity, where the same content is presented as a real and virtual source in short succession [48].

During the experiment participants were asked to walk along a predefined path in the listening area. According to the test paradigm, a random test signal was played from one of the four sources (real or virtual) and participants had to answer whether the presentation was real or virtual by pressing the corresponding button on the wireless MIDI controller. The participants were allowed to walk the path at their desired speed, turn around, or stop for a short time. In general, however, they were asked to cover the entire listening area by constantly walking along the path.

3.2.1 Results

Data collection was restricted due to the effects of the COVID-19 pandemic, resulting in a pilot study with four participants (1 female, 3 male, aged 25 to 35 years). All participants were highly experienced with binaural reproduction since they work full-time in this field and they were not naive as to the purpose of this study nor the technical background of the 6-DoF VAE.

The participants had a detection rate of $p_{2\mathrm{AFC}} = 0.61$, 0.77, 0.84, and 0.87, corresponding to $d' = 0.40$, 1.04, 1.41, and 1.59. Concerning the critical value of $d' = 0.82$, the presented 6-DoF parametric spatial audio system was thus perceptually plausible (in this strictly defined manner) for 1 participant, whereas the other participants could more reliably detect whether a real or virtual sound source was presented.

There are several reasons why the participants could distinguish between virtual and real sources, as revealed by informal discussions after the experiments. First of all it is important to mention again that the pilot study participants are to be classified as extremely critical listeners since they are experts in this field and have also followed the development of the system. As typical indicators for detecting a virtual source the participants mentioned the apparent source width, the listener envelopment, and slight localization inaccuracies between a virtual source and physically

present loudspeaker, especially at front grid positions very close to speaker 1 or 4.

These observations indicate that the BRIR synthesis has to be further improved, especially in terms of spaciousness (IC and IACC), and that the spatial resolution of 4° in azimuth and 10° in elevation, which we have chosen as a compromise between accuracy and memory capacities, might be too low. The greater apparent source width of the virtual sources sometimes perceived by the participants as well as localization errors may also have been caused by the employed non-individual KEMAR HRTFs.

However, regardless of the quality of the synthesized BRIRs, there were several system-related effects that allowed participants to distinguish between a real and virtual source. Two of the four participants used head-tracking latency during high-speed rotations, which caused the virtual sound source to lag behind, to detect virtual sources. Furthermore, especially with periodic continuous test signals like, for example, legato strings, participants could perceive the BRIR switching at grid transitions, i.e., when walking through the room and passing various grid cells. These findings are in line with a study from Werner et al. [49], which showed that especially at grid transitions (close to the sound source), participants perceived abnormalities in the synthesis. The artifact was more pronounced with audio content such as strings since the different BRIRs are cross-faded at the grid transitions, which leads to audible phase shifts, especially with periodic content. Speech or drum test signals therefore did not lead to such audible artifacts at grid transitions, which is also in line with the results from Werner et al. [49].

Furthermore despite moderate playback levels, the participants could sometimes feel the headphones' vibration when a binaural signal was presented and could thus detect whether the source was real or virtual. This problem could easily be solved with other more extraaural headphones. Lastly we could observe an apparent training effect, that is, the participants got to know all these artifacts throughout the experiment and then recognized the virtual source often due to the artifacts but not necessarily due to audible differences to a real source.

4 DISCUSSION

Parametric rendering is a promising approach to create perceptually plausible spatial audio renderings of high quality for applications with low computing capacity, such as AR applications. Most parametric encoding methods rely on SRIR measurements, such as first-order Ambisonics (FOA) or open array measurements [1, 3, 2, 5, 6], or even higher-order rigid-sphere array measurement [4, 25]. Those methods usually auralize the sound field only at the measurement point for the listener's head orientation, i.e., the sound field is only rendered for 3 DoF.

More recently proposed methods for 6-DoF rendering typically require at least one [7, 8] or even multiple SRIRs [9, 10] from different positions in the room and apply extrapolation based on one SRIR or interpolation between the distributed SRIRs to incorporate listener translation. However 6-DoF rendering based on a parametric description of the sound field at one position in the space, encoded from a single RIR measurement, is rare. A perceptually plausible parametric 6-DoF rendering based on one RIR seems to be a good solution, though, e.g., for AR-glasses to augment a real environment with virtual sources, taking into account the listener's movements.

With the *Paraspax* method presented in this paper we have taken a first step toward scalable and perceptually motivated parametric 6-DoF rendering based on the encoding of a single monaural RIR. The three basic processing blocks for encoding include the parametrization of the monaural RIR to determine basic monaural room acoustic parameters and TOAs of the direct sound and early reflections, the spatialization to assign DOAs to the direct sound and the selected reflections, and the optional extrapolation for listener translation. The estimated parameters can then be used for synthesizing BRIRs based on the monaural RIR, as presented in this paper, or for real-time parametric rendering (see, e.g., [29, 50]). As the encoding provides the (perceptually) most relevant parameters to describe the sound field, the real-time rendering could be performed without any signal components of the monaural RIR using computationally more efficient auralization methods than the real-time convolution with a large number of precomputed BRIRs.

The presented method adapts some ideas from the implementation of Pörschmann et al. [11], who published a first approach to synthesize BRIRs that are generally suitable for 6-DoF rendering based on an omnidirectional RIR. In comparison, however, the *Paraspax* method provides a scalable and perceptually motivated reflection detection, i.e., the number of salient reflections [12] that are en and de-coded for rendering can be adjusted according to the available computing resources. Furthermore the three different options for spatialization allow estimating DOAs without any spatial information (pseudo-randomized), based on few geometric data (image source model), or based on SRIR measurements or simulations (precomputed). Thus even if only monaural parameters are available the *Paraspax* method allows 6-DoF parametric spatial audio rendering.

Of course spatial percepts such as apparent source width, localizability, or listener envelopment can be impaired by only partially correct early reflection DOAs, but listening experiments also showed a high perceptual similarity between measured reference BRIRs and synthesized BRIRs with pseudo-randomized DOAs [11, 23], suggesting that in cases where there is no geometric or spatial information at all, such an auralization may still be a sufficiently good solution. On the other hand passing precomputed DOAs allows combining the *Paraspax* en and de-coding with correctly estimated DOAs. Using image source simulation for spatialization provides DOA estimates for which the accuracy depends on the room geometry and which therefore become less accurate the more the actual room geometry deviates from the assumed shoebox geometry.

Informal listening for the present room showed that spatialization with pseudo-random DOAs produced audible differences, especially in terms of listener envelopment

and spaciousness, compared to the other two spatialization methods (ISM and precomputed DOAs from SDM measurements), which could not be perceptually distinguished. However since the examined room is very close to the shape of a shoebox we did not expect significant perceptual differences between these two spatialization methods.

The *Paraspax* method also includes the sound source directivity in the BRIR synthesis. Especially due to the emerging research in 6-DoF audio reproduction, the topic of source directivity has gained more attention again [51, 52]. In particular scenarios where the listener can walk around a virtual sound source should benefit from considering the directivity, but even if the sources are outside the listening area, incorporating the directivity leads to better technical and perceptual results [51].

In the present plausibility experiment, however, participants could not walk around the loudspeaker, limiting our results concerning the benefit of including source directivity in the BRIR synthesis. Even though directivity effects are also audible when the loudspeakers are outside the listening area, especially at lateral positions very close to the loudspeaker, scenarios in which the listener can walk around the loudspeaker reveal effects of the directivity much more strongly and thus also allow a better evaluation of including source directivity, for example, concerning plausibility. Besides, further research is required to examine how accurately the directivity needs to be reproduced and how precisely it needs to be integrated into the synthesis (i.e., whether for example the early reflections also need to be adjusted according to the source directivity).

The physical evaluation presented in SEC. 2 revealed a good performance of the BRIR synthesis. At most of the tested positions in the room differences in DRR, RT_{30}, and IACC to reference measurements were below the respective JND, indicating inaudible differences, at least concerning the examined parameters. We obtained similarly good results with the three other sources in the same room and also tested the encoding with various RIRs of different rooms. However presenting all these results would go beyond the scope of this paper, so we decided to show a detailed physical evaluation based on the room in which we set up the plausibility experiment.

Overall the results suggest that the encoding, synthesis, and extrapolation work correctly. However the presented extrapolation treats early reflections as image sources and adjusts amplitude, TOAs, and DOAs accordingly when the listener moves through the room. Strictly speaking this is only correct if a) the selected early reflections originate from the surfaces of a shoebox room and not, for example, from reflecting objects in the room and b) the DOAs were correctly assigned to the selected reflections in the first place. Furthermore such an extrapolation of parameters is not necessarily correct in rooms with complex geometries. Therefore more research in different, more complex room geometries is required to further evaluate the extrapolation method and optimize the procedure for more general applicability.

In AR applications however the direction and distance of the direct sound are either predefined or can be estimated by using multimodal information, e.g., RGB or depth cameras, Simultaneous Localization and Mapping (SLAM), and can thus always be correctly adjusted according to the listener's movements and head orientation. Thus an incorrect extrapolation model would only negatively affect the early reflections. Additionally multimodal information could also be used to obtain rough estimations of room geometry and dimensions to inform the DOA generation process. However how precisely the early reflection pattern has to be adjusted according to the listener's movements to achieve a plausible reproduction and whether it always has to be dynamically adjusted at all needs further research.

The real-time framework presented in SEC. 3 allows experiencing a VAE with 6 DoF. The listener can move through the room, switching between real and virtual sources or changing the test signal, which allows an intuitive evaluation of the VAE. The system is not tied to the use of the *Paraspax* BRIR synthesis, i.e., any BRIRs can be used, and it can be flexibly extended, e.g., to implement real-time parametric rendering instead of convolution-based synthesis.

Based on the real-time framework we implemented a plausibility experiment according to Lindau and Weinzierl [40], which can be regarded as the strictest form of plausibility testing, as it is based on a comparison with real sound sources. According to this strict interpretation the 6-DoF VAE was plausible for one of the four expert listeners participating in the pilot study. Another participant had a detection rate only slightly above the critical value, whereas the other two participants could distinguish more reliably between virtual and real sources.

However due to the small number of participants and the fact that they were expert listeners we cannot yet draw any final conclusions, and we plan in future work to evaluate the perceptual plausibility of the presented 6-DoF VAE with a higher number of naive listeners. In particular we plan to examine the plausibility of the VAE for 6-DoF scenarios with different grid resolutions and trajectories, but we also plan to conduct 3-DoF experiments in which participants stand at different positions in the room and evaluate the plausibility of the system. Comparing those conditions allows us to investigate, for example, to what extent movement in a VAE can affect the perceived plausibility.

Interestingly the interviews with the participants of the pilot study revealed that often technical artifacts of the system made it possible to distinguish between a real and virtual source and not necessarily perceptible differences related to the synthesized BRIRs. Thus the pilot study showed that in order to achieve this strict form of plausibility, in which participants always compare with real sources, not only a high-quality BRIR synthesis is required but also a flawless technical framework, ideally without any artifacts that can be used by the listener to detect a virtual source. However, despite the artifacts described, the auralization is still very convincing and we therefore hypothesize that after solving the discussed technical challenges and further optimizing the real-time system, a plausible 6-DoF VAE based on one monaural RIR, tested as strictly as described,

is feasible. To get an impression of the quality of the BRIR synthesis we provide some audio examples online.[2]

5 CONCLUSION

In this paper we presented a method for 6-DoF parametric spatial audio reproduction based on one monaural RIR. The *Paraspax* toolbox provides the entire pipeline to derive a spatial parametric description of the sound field from a monaural RIR and generate synthetic BRIRs for any desired head orientation and position in the room. The physical evaluation showed a good performance of the method, mostly with differences to reference measurements below the JND of the considered room acoustic parameter at all tested positions in the room. As a basis for a 6-DoF VAE we implemented a real-time framework that uses the synthesized BRIRs and enables both demo applications and psychoacoustic experiments in 6-DoF environments.

Using the framework we carried out a pilot study with expert listeners to assess the plausibility of the 6-DoF VAE. The results showed that the 6-DoF VAE in its current state can provide a plausible binaural reproduction for a listener moving through the room. However the results also revealed specific technical challenges for 6-DoF systems producing artifacts that are not directly related to the quality of the BRIR synthesis or the auralization but make it easier for participants to distinguish between real and virtual sources. Therefore our future work will mainly focus on optimizing the real-time framework to further enhance the plausibility of the 6-DoF VAE, as well as conducting listening experiments with naive listeners.

6 REFERENCES

[1] J. Merimaa and V. Pulkki, "Spatial Impulse Response Rendering I: Analysis and Synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127 (2005 Dec.).

[2] V. Pulkki, "Spatial Sound Reproduction With Directional Audio Coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516 (2007 Jun.).

[3] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial Decomposition Method for Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 17–28 (2013 Jan.).

[4] P. Stade, J. M. Arend, and C. Pörschmann, "Perceptual Evaluation of Synthetic Early Binaural Room Impulse Responses Based on a Parametric Model," presented at the *142nd Convention of the Audio Engineering Society*, pp. 1–10 (2017 May), paper 9688.

[5] P. Coleman, A. Franck, D. Menzies, and P. J. B. Jackson, "Object-Based Reverberation Encoding From First-Order Ambisonic RIRs," presented at the *142nd Convention of the Audio Engineering Society*, pp. 1–10 (2017 May), paper 9731.

[6] M. Zaunschirm, M. Frank, and F. Zotter, "BRIR Synthesis Using First-Order Microphone Arrays," presented at the *144th Convention of the Audio Engineering Society*, pp. 1–10 (2018 May), paper 9944.

[7] T. Pihlajamäki and V. Pulkki, "Synthesis of Complex Sound Scenes With Transformation of Recorded Spatial Sound in Virtual Reality," *J. Audio Eng. Soc.*, vol. 63, no. 7/8, pp. 542–551 (2015 Jul.). http://dx.doi.org/10.17743/jaes.2015.0059.

[8] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. A. P. Habets, "Six-Degrees-of-Freedom Binaural Audio Reproduction of First-Order Ambisonics With Distance Information," in *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality*, pp. 1–10 (2018 Aug.), paper P6-2.

[9] J. G. Tylka and E. Y. Choueiri, "Domains of Practical Applicability for Parametric Interpolation Methods for Virtual Sound Field Navigation," *J. Audio Eng. Soc.*, vol. 67, no. 11, pp. 882–893 (2019 Nov.). https://doi.org/10.17743/jaes.2019.0038.

[10] K. Müller and F. Zotter, "Auralization Based on Multi-Perspective Ambisonic Room Impulse Responses," *Acta Acust.*, vol. 4, no. 6, pp. 1–18 (2020 Nov.). https://doi.org/10.1051/aacus/2020024.

[11] C. Pörschmann, P. Stade, and J. M. Arend, "Binauralization of Omnidirectional Room Impulse Responses - Algorithm and Technical Evaluation," in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*, pp. 345–352 (Edinburgh, UK) (2017 Sep.).

[12] F. Brinkmann, H. Gamper, N. Raghuvanshi, and I. Tashev, "Towards Encoding Perceptually Salient Early Reflections for Parametric Spatial Audio Rendering," presented at the *148th Convention of the Audio Engineering Society*, pp. 1–11 (2020 May), paper 10380.

[13] J. S. Abel and P. Huang, "A Simple, Robust Measure of Reverberation Echo Density," presented at the *121st Convention of the Audio Engineering Society*, pp. 1–10 (2006 Oct.), paper 6985.

[14] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1996).

[15] G. Defrance, L. Daudet, and J.-D. Polack, "Finding the Onset of a Room Impulse Response: Straightforward?" *J. Acoust. Soc. Am.*, vol. 124, no. 4, pp. EL248–EL254 (2008 Oct.). https://doi.org/10.1121/1.2960935.

[16] S. Bech, "Timbral Aspects of Reproduced Sound in Small Rooms. I," *J. Acoust. Soc. Am.*, vol. 97, no. 3, pp. 1717–1726 (1995 Mar.).

[17] B. Gourévitch and R. Brette, "The Impact of Early Reflections on Binaural Cues," *J. Acoust. Soc. Am.*, vol. 132, no. 1, pp. 9–27 (2012 Jul.). https://doi.org/10.1121/1.4726052.

[18] S. E. Olive and F. E. Toole, "The Detection of Reflections in Typical Rooms," presented at the *85th Convention of the Audio Engineering Society*, pp. 1–36 (1988 Nov.), paper 2719.

[19] P. Coleman, A. Franck, P. J. B. Jackson, R. J. Hughes, L. Remaggi, and F. Melchior, "Object-Based Reverberation for Spatial Audio," *J. Audio Eng. Soc.*, vol. 65,

[2] Available: https://github.com/facebookresearch/Paraspax/tree/main/Paraspax_AudioExamples

no. 1/2, pp. 66–77 (2017 Jan.). https://doi.org/10.17743/jaes.2016.0059.

[20] E. Brandão, A. Lenzi, and S. Paul, "A Review of the *In Situ* Impedance and Sound Absorption Measurement Techniques," *Acta Acust. United Acust.*, vol. 101, no. 3, pp. 443–463 (2015 May/Jun.). https://doi.org/10.3813/AAA.918840.

[21] B. Friedlander and B. Porat, "The Modified Yule-Walker Method of ARMA Spectral Estimation," *IEEE Trans. Aero. Electr. Syst.*, vol. AES-20, no. 2, pp. 158–173 (1984 Mar.). https://doi.org/10.1109/TAES.1984.310437.

[22] J. M. Arend, T. Lübeck, and C. Pörschmann, "A Reactive Virtual Acoustic Environment for Interactive Immersive Audio," in *Proceedings of the AES International Conference on Immersive and Interactive Audio*, pp. 1–10 (2019 Mar.), paper 9.

[23] J. Ahrens, "Auralization of Omnidirectional Room Impulse Responses Based on the Spatial Decomposition Method and Synthetic Spatial Data," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 146–150 (Brighton, UK) (2019 May).

[24] 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, no. 11, pp. 914–925 (2015 Nov.). https://doi.org/10.17743/jaes.2015.0080.

[25] 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, no. 5, pp. 338–354 (2020 May). https://doi.org/10.17743/jaes.2020.0026.

[26] S. V. Amengual Garí, J. M. Arend, P. T. Calamia, and P. W. Robinson, "Optimizations of the Spatial Decomposition Method for Binaural Reproduction," *J. Audio Eng. Soc.*, vol. 68, no. 12, pp. 959–976 (2020 Dec.). https://doi.org/10.17743/jaes.2020.0063.

[27] A. J. Kolarik, B. C. J. Moore, P. Zahorik, S. Cirstea, and S. Pardhan, "Auditory Distance Perception in Humans: A Review of Cues, Development, Neuronal Bases, and Effects of Sensory Loss," *Atten. Percept. Psych.*, vol. 78, no. 2, pp. 373–395 (2016 Feb.). https://doi.org/10.3758/s13414-015-1015-1.

[28] F. Menzer and C. Faller, "Investigations on an Early-Reflection-Free Model for BRIRs," *J. Audio Eng. Soc.*, vol. 58, no. 9, pp. 709–723 (2010 Sep.).

[29] T. Wendt, S. van de Par, and S. D. Ewert, "A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation," *J. Audio Eng. Soc.*, vol. 62, no. 11, pp. 748–766 (2014 Nov.). https://doi.org/10.17743/jaes.2014.0042.

[30] G. S. Kendall, "The Decorrelation of Audio Signals and Its Impact on Spatial Imagery," *Comp. Music J.*, vol. 19, no. 4, pp. 71–87 (1995). https://doi.org/10.2307/3680992.

[31] J. Blauert, *The Technology of Binaural Listening* (Springer-Verlag Berlin Heidelberg, Heidelberg, Germany, 2013).

[32] C. Borß and R. Martin, "An Improved Parametric Model for Perception-Based Design of Virtual Acoustics," in *Proceedings of the 35th International Conference: Audio for Games*, pp. 1–8 (2009 Feb.), paper 3.

[33] H. Kuttruff, *Room Acoustics (Fifth Edition)* (CRC Press, Boca Raton, FL, 2009).

[34] B. Rafaely, *Fundamentals of Spherical Array Processing* (Springer-Verlag Berlin Heidelberg, Heidelberg, Germany, 2015). https://doi.org/10.1007/978-3-662-45664-4.

[35] B. Xie, "On the Low Frequency Characteristics of Head-Related Transfer Function," *Chinese J. Acoust.*, vol. 28, no. 2, pp. 116–128 (2009).

[36] L. Aspöck, F. Brinkmann, D. Ackermann, S. Weinzierl, and M. Vorländer, "BRAS - Benchmark for Room Acoustical Simulation," http://dx.doi.org/10.14279/depositonce-6726.2 (2019).

[37] S. Csadi, F. M. Boland, L. Ferguson, H. O'Dwyer, and E. Bates, "Direct to Reverberant Ratio Measurements in Small and Mid-Sized Rooms," in *Proceedings of the AES International Conference on Immersive and Interactive Audio*, pp. 1–10 (2019 Mar.), paper 28.

[38] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the Minimum Audible Difference in Direct-to-Reverberant Energy Ratio," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 450–461 (2008 Jul.). https://doi.org/10.1121/1.2936368.

[39] H. Seraphim, "Untersuchungen über die Unterschiedsschwelle exponentiellen Abklingens von Rauschbandimpulsen," *Acta Acust. United Acust.*, vol. 8, supp. 1, pp. 280–284 (1958).

[40] A. Lindau and S. Weinzierl, "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acust. United Acust.*, vol. 98, no. 5, pp. 804–810 (2012 Sep./Oct.). https://doi.org/10.3813/AAA.918562.

[41] A. Neidhardt, F. Klein, N. Knoop, and T. Köllmer, "Flexible Python Tool for Dynamic Binaural Synthesis Applications," presented at the *142nd Convention of the Audio Engineering Society*, pp. 1–5 (2017 May), paper 346.

[42] J. Gómez-Bolaños, A. Mäkivirta, and V. Pulkki, "Automatic Regularization Parameter for Headphone Transfer Function Inversion," *J. Audio Eng. Soc.*, vol. 64, no. 10, pp. 752–761 (2016 Oct.). http://dx.doi.org/10.17743/jaes.2016.0030.

[43] A. Lindau, "The Perception of System Latency in Dynamic Binaural Synthesis," presented at the *35th DAGA*, pp. 1063–1066 (2009 Mar.).

[44] T. Waltemate, F. Hülsmann, T. Pfeiffer, S. Kopp, and M. Botsch, "Realizing a Low-Latency Virtual Reality Environment for Motor Learning," in *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology*, pp. 139–147 (2015 Nov.). https://doi.org/10.1145/2821592.2821607.

[45] T. Robotham, O. Rummukainen, J. Herre, and E. A. P. Habets, "Evaluation of Binaural Renderers in Virtual

Reality Environments: Platform and Examples," presented at the *145th Convention of the Audio Engineering Society*, pp. 1–5 (2018 Oct.), paper 454.

[46] S. V. Amengual Garí, W. O. Brimijoin, H. G. Hassager, and P. W. Robinson, "Flexible Binaural Resynthesis of Room Impulse Responses for Augmented Reality Research," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, pp. 161–166 (Paris, France) (2019 Sep.). https://doi.org/10.25836/sasp.2019.31.

[47] 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, no. 4, pp. 2746–2760 (2019 Apr.). https://doi.org/10.1121/1.5096178.

[48] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the Authenticity of Individual Dynamic Binaural Synthesis," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 1784–1795 (2017 Oct.). https://doi.org/10.1121/1.5005606.

[49] S. Werner, F. Klein, and G. Götz, "Investigation on Spatial Auditory Perception using Non-Uniform Spatial Distribution of Binaural Room Impulse Responses," in *Proceedings of the 5th International Conference on Spatial Audio (ICSA)*, pp. 137–144 (2019 Sep.). https://doi.org/10.22032/dbt.39967.

[50] C. Schissler, P. Stirling, and R. Mehra, "Efficient Construction of the Spatial Room Impulse Response," in *Proceedings of the IEEE Virtual Reality (VR)*, pp. 122–130 (Los Angeles, CA) (2017 Mar.). https://doi.org/10.1109/VR.2017.7892239.

[51] U. Sloma, F. Klein, S. Werner, and T. Pappachan Kannookadan, "Synthesis of Binaural Room Impulse Responses for Different Listening Positions Considering the Source Directivity," presented at the *147th Convention of the Audio Engineering Society*, pp. 1–9 (2019 Oct.), paper 10237.

[52] T. Robotham, O. S. Rummukainen, and E. A. P. Habets, "Towards the Perception of Sound Source Directivity Inside Six-Degrees-of-Freedom Virtual Reality," in *Proceedings of the 5th International Conference on Spatial Audio (ICSA)*, pp. 71–78 (Ilmenau, Germany) (2019 Sep.). https://doi.org/10.22032/dbt.39956.

APPENDIX

Table 1. Estimated DRR, RT_{30}, and IACC values for the measured RIRs (RIR_M), measured BRIRs ($BRIR_M$), and synthesized BRIRs ($BRIR_S$) at the grid positions.

Position	DRR in dB			RT_{30} in s			IACC	
	RIR_M	$BRIR_M$	$BRIR_S$	RIR_M	$BRIR_M$	$BRIR_S$	$BRIR_M$	$BRIR_S$
1	−4.73	−3.10	−3.73	0.90	0.85	0.92	0.21	0.14
2	−2.59	···	−2.89	0.88	···	0.91	···	0.38
3	−2.62	−1.86	−2.58	0.90	0.84	0.91	0.40	0.43
4	−4.48	···	−2.97	0.91	···	0.91	···	0.25
5	−6.56	···	−5.08	0.88	···	0.92	···	0.19
6	−5.18	−3.72	−4.61	0.93	0.84	0.92	0.31	0.23
7	−4.83	−4.16	−4.33	0.90	0.83	0.92	0.34	0.36
8	−6.10	···	−4.42	0.91	···	0.92	···	0.24
9	−8.27	−6.47	−6.48	0.88	0.82	0.91	0.18	0.18
10	−7.14	−5.89	−6.05	0.90	0.84	0.92	0.29	0.35
11	−7.04	−6.00	−5.89	0.90	0.82	0.92	0.26	0.35
12	−7.61	−6.61	−5.92	0.90	0.84	0.91	0.16	0.24
13	−9.49	···	−7.25	0.89	···	0.92	···	0.17
14	−8.78	···	−7.21	0.91	···	0.91	···	0.22
15	−8.90	−7.76	−7.02	0.92	0.83	0.92	0.24	0.32
16	−9.20	−7.95	−6.87	0.89	0.83	0.92	0.17	0.24
17	−10.74	···	−8.15	0.87	···	0.92	···	0.20
18	−9.84	−8.26	−8.13	0.92	0.85	0.92	0.26	0.32
19	−10.09	−9.06	−8.18	0.89	0.83	0.92	0.18	0.32
20	−10.25	···	−7.73	0.89	···	0.92	···	0.23

DRR calculated with a direct sound window of 1.5 ms. For BRIRs, the direct and diffuse sound energy of both channels were summed before DRR estimation. RT_{30} calculated as the mean in the octave bands from 0.25–8 kHz. For BRIRs, determined broadband RT_{30} values per channel were averaged to obtain a single value. Missing values due to missing measurement data indicated by (···).

THE AUTHORS


Johannes M. Arend


Sebastià V. Amengual Garí


Carl Schissler


Florian Klein


Philip W. Robinson

Johannes M. Arend received a B.Eng. degree in media technology from HS Düsseldorf (Germany) in 2011 and an M.Sc. degree in media technology from TH Köln (Germany) in 2014. Since 2015 he has been a Research Fellow and working toward a Ph.D. degree at TH Köln and TU Berlin in the field of spatial audio with a focus on binaural technology, auditory perception, and audio signal processing. Between September 2019 and March 2020 he was a research intern at Facebook Reality Labs Research.

•

Sebastià V. Amengual Garí is currently a research scientist at Facebook Reality Labs Research working on room acoustics, spatial audio, and auditory perception. He received a Diploma Degree in Telecommunications with a major in Sound and Image in 2014 from the Polytechnic University of Catalonia (UPC) in 2014, completing his Master's Thesis at the Norwegian University of Science and Technology (NTNU). His doctoral work at the Detmold University of Music focused on investigating the interaction of room acoustics and live music performance using virtual acoustic environments. His research interests lie in the intersection of audio, perception, and music.

•

Carl Schissler is currently a research scientist at Facebook Reality Labs Research, which he joined in 2017 after receiving his Ph.D. in computer science from the University of North Carolina at Chapel Hill. His primary research interests include real-time acoustic simulation, digital signal processing, and computational geometry. Beyond research, Carl is a multi-instrumentalist and amateur composer and enjoys audio mixing for recording and stage.

•

Florian Klein is a Ph.D. student at Technical University Ilmenau, Germany. His main research area is auditory adaptation processes in the scope of spatial hearing. Furthermore he is working on solutions for 6DOF spatial sound rendering for AR/VR applications. He received the Best Student Paper Award at the AES Convention 2015 and a Best Paper Award at the AES Convention 2019. Between November 2019 and May 2020 he was a research intern at Facebook Reality Labs Research.

•

Philip W. Robinson is a research science manager in room acoustics and auditory perception at Facebook Reality Labs Research (FRL Research) in Redmond, WA. Prior to joining FRL Research he incorporated virtual acoustics simulation and reproduction systems into building design processes at the architecture firm of Foster + Partners. He was a Fulbright Scholar and post-doctoral researcher at Aalto University in Finland, where he studied perception of concert hall acoustics, spatial auditory resolution, and echo thresholds. He has been a visiting researcher at EPFL in Switzerland and Hanyang University in South Korea. He received a Ph.D. from Rensselaer Polytechnic Institute in Troy, NY in 2012. In a previous life he was a registered architect in his home state of New Mexico. He remains passionate about architecture, the study of which gave him a great interest in perception of environments, real or virtual.

3.2 Optimizations of the Spatial Decomposition Method for Binaural Reproduction

Amengual Garí, S. V., Arend, J. M., Calamia, P., & Robinson, P. W. (2020). *J. Audio Eng. Soc.*, *68*(12), 959–976. https://doi.org/10.17743/jaes.2020.0063

OPEN ACCESS Freely available online
S. V. A. Garí, J. M. Arend, P. T. Calamia and P. W. Robinson, "Optimizations of the Spatial Decomposition Method for Binaural Reproduction"
J. Audio Eng. Soc., vol. 68, no. 12, pp. 959–976, (2020 December).
DOI: https://doi.org/10.17743/jaes.2020.0063

Optimizations of the Spatial Decomposition Method for Binaural Reproduction

SEBASTIÀ V. AMENGUAL GARÍ,[1] *AES Member*, **JOHANNES M. AREND,**[1,2] *AES Student Member*,
(samengual@fb.com) (johannes.arend@th-koeln.de)

PAUL T. CALAMIA,[1] **AND PHILIP W. ROBINSON,**[1] *AES Member*
(pcalamia@fb.com) (philrob22@fb.com)

[1] *Facebook Reality Labs Research, Redmond, WA*
[2] *TH Köln - University of Applied Sciences, Cologne, Germany*

The spatial decomposition method (SDM) can be used to parameterize and reproduce a sound field based on measured multichannel room impulse responses (RIRs). In this paper we propose optimizations of SDM to address the following questions and issues that have recently emerged in the development of the method: (a) accuracy in direction-of-arrival (DOA) estimation with open microphone arrays utilizing time differences of arrival as well as with B-format arrays using pseudo-intensity vectors; (b) optimal array size and temporal processing window size for broadband DOA estimation based on open microphone arrays; (c) spatial and spectral distortion of single events caused by unstable DOA estimation; and (d) spectral whitening of late reverberation as a consequence of rapidly varying DOA estimates. Through simulations we analyze DOA estimation accuracy (a) and explore processing parameters (b) in search of optimal settings. To overcome the unnatural DOA spread (c), we introduce spatial quantization of the DOA as a post-processing step at the expense of spatial distortion for successive reflections. To address the spectral whitening (d), we propose an equalization approach specifically designed for rendering SDM data directly to binaural signals with a spatially dense HRTF dataset. Finally, through perceptual experiments, we evaluate the proposed equalization and investigate the consequences of quantizing the spatial information of SDM auralizations by directly comparing binaural renderings with real loudspeakers. The proposed improvements for binaural rendering are released in an open source repository.*

0 INTRODUCTION

The spatial decomposition method (SDM) [1] can be used to parameterize and reproduce a sound field based on measured multichannel room impulse responses (RIRs). The direction of arrival (DOA) of each sample of the RIR is first estimated and then the instantaneous energy is mapped to its corresponding direction using any loudspeaker or headphone-based reproduction method.

Since its initial conception the method has been used for multiple applications, such as concert hall analysis and auralization [2–4], stage acoustics [5, 6], car cabin acoustics [7, 8], acoustic preference research in small rooms [9, 10], speech intelligibility [11], or audio-visual perception in virtual reality (VR) [12]. Additionally it has been used in conjunction with multiple spatial audio reproduction methods, such as Vector-Base Amplitude Panning (VBAP) [13, 14], Nearest Loudspeaker Synthesis (NLS) [7, 15, 4], Ambisonics [16, 17], and binaural synthesis [16, 6]. The increasing popularity of the method can be partially attributed to the publicly available SDM Toolbox [18], released by the original developers of the method.

The extended usage of the method resulted in a series of questions and open issues. In this manuscript we focus on the case of the analysis of sound fields composed of broadband single sources in small and medium rooms and reproduced via binaural synthesis. In particular we address four particular topics in this work: (a) SDM can be used with various DOA estimation methods, such as time differences of arrival (TDOA) using open microphone arrays [1] or pseudo-intensity vectors (PIV) using B-format arrays [7].

*https://github.com/facebookresearch/BinauralSDM

However it is not yet clear which of the two estimation methods provides more accurate DOAs and under which conditions one of the two methods performs better.

(b) Various parameters such as array size and temporal processing window size for DOA estimation with an open microphone array have not been investigated systematically, and the question remains as to what are the optimal parameters for best possible DOA detection using broadband RIRs. (c) As the RIR progresses into the late reverberation, the DOA estimation becomes unreliable [1], leading to spatial spread of single events—a problem not yet addressed when using SDM for binaural reproduction.

(d) As a result of rapidly varying DOA estimates in the late part of the RIR, the late reverberation becomes spectrally white [7, 17]. However current solutions to overcome this issue [7] are computationally inefficient when rendering SDM data directly to binaural signals, requiring a suitable alternative for binaural reproduction. In the following paragraphs we introduce the above-mentioned points in more detail and outline how we examine the technical questions. We then briefly describe our proposed optimizations of SDM for the generation of binaural renderings, covering the three stages of processing—measurement, analysis, and rendering.

The original SDM method, developed for open microphone arrays and performing the DOA analysis using TDOA, was validated numerically and perceptually with reference to an image source model [1]. Later the same authors released an open implementation of the algorithm [18], including an alternative analysis approach based on broadband PIVs of B-format RIRs (similar to [19]), which has lately been popularized, and enabling the usage of the method with a greater variety of array configurations. Recent evaluations suggest that the DOAs are not reliably estimated when analyzing broadband B-format signals [20], although perceptually satisfactory results can be obtained with appropriate bandpass filtering of the raw RIRs and subsequent smoothing of the DOA estimates [16].

In this paper, we explore the analysis requirements and compare simulation and measurement results from PIV analysis to those of TDOA. Furthermore we investigate optimal parameters for the analysis using TDOA and open arrays. Note that we focus our investigations on the analysis of broadband events. Analysis using multiple bands, such as in [7], results in additional degrees of freedom in the search for optimal parameter values, which could be different in each analysis band.

In regard to headphone reproductions of SDM RIRs, the use of dense Head-Related Transfer Function (HRTF) datasets results in a higher degree of spatial resolution at the cost of potential timbral degradations. As the RIR progresses into the late reverberation, the DOA estimates become unstable and less reliable [1]. This causes consecutive samples of the RIR to be mapped to disparate locations—an effect that is accentuated by the fact that small fluctuations will result in reflections being mapped onto several adjacent HRTFs. To address this we discuss approaches for the post-processing of DOAs based on the spatial quantization and clustering of reflections, reducing the DOA spread significantly at the expense of clustering consecutive reflections onto the same directions. In particular we focus on the implications of using regular grids for quantization of the spatial information.

Rapidly varying DOA estimates cause a spectral whitening in certain portions of the rendered responses, as consecutive samples corresponding to the same band-limited event are mapped onto disparate locations as broadband events. This is especially relevant in small spaces with high reflection density [7] or at the late reverberation tail [16, 5, 6], where the DOA cannot be reliably estimated. The presence of this artifact has been reported with multiple reproduction approaches, such as NLS [7], Ambisonics [17], or binaural synthesis [6], and in typical rooms it generally results in an increase of the reverberation time at high frequencies.

Tervo et al. proposed a time-frequency equalization to address this problem and validated it in the application of car cabin acoustics [7]. This equalization method was designed for loudspeaker reproduction, and it generates time-varying filters for each of the loudspeaker (rendered) RIRs by comparing the average magnitude response of the rendered RIRs and original pressure RIR. Applying this approach to binaural rendering is possible by using a virtual loudspeaker approach, where each loudspeaker feed is convolved with the Head-Related Impulse Response (HRIR) corresponding to the loudspeaker location. However, with a spatially dense HRTF dataset, this approach becomes impractical due to computing limitations. In this paper we propose an alternative equalization approach comprising a reverberation correction process (RTMod) and the processing of the resulting Binaural Room Impulse Responses (BRIR) with a cascade of allpass filters (RTMod+AP).

Finally, as has been suggested previously [21], we hypothesize that the spatial resolution of the SDM auralizations can be reduced without perceivable degradations. We investigate the minimum required spatial resolution in perceptual experiments employing SDM auralizations by directly comparing binaural renderings with real loudspeakers.

The paper is structured as follows. Sec. 1 reviews the two approaches (TDOA and PIV) used in SDM for the DOA analysis. Sec. 2 evaluates the performance of the directional analysis for various common array and parameter configurations using simulations. Sec. 3 compares the results of the directional analysis conducted with TDOA and PIV on the same set of measurements from a tetrahedral array. Sec. 4 describes our proposed rendering approach to resynthesize binaural RIRs, including DOA post-processing, a novel equalization method for the reverberation and instrumental validation. Sec. 5 presents a perceptual evaluation on the plausibility of BRIRs with quantized spatial resolution. Secs. 6 and 7 present a discussion and conclusions, respectively.

1 DOA ESTIMATION

The basic paradigm of SDM involves assigning one DOA to each of the samples of a pressure RIR, implicitly assuming that the sound field is composed of a succession of

broadband specular events. Once this information is available it can be used for the directional analysis of an RIR or to re-synthesize the sound field using any loudspeaker or headphone-based method. Two main approaches are currently widespread to perform the DOA analysis, depending on the nature of the microphone array and available signals.

1.1 Time Differences of Arrival (TDOA) Method

In this section we review the method introduced by Tervo et al. in [1], which estimates DOA data from a multichannel RIR by exploiting the TDOAs between microphones. The estimation requires an open array of $M \geq 4$ microphones arranged in a 3D space. Although the authors recommended the use of omnidirectional microphones, accurate results have been obtained with arrays of cardioid microphones as well [15], suggesting that the requirements regarding directivity are somewhat flexible. However, if the data are intended to be used for auralization, at least one of the microphones must be omnidirectional or encoded to present an omnidirectional response. Alternatives to encode directional responses are proposed in [7] but are beyond the scope of this paper.

A sliding Hanning window of size L is applied to the RIR and at each time step the DOA is resolved for one single acoustic event. The window is moved in 1-sample steps and thus one DOA is estimated for each sample of the RIR. The size of the sliding window must be equal to or greater than the time needed for a plane wave to travel between the most distant microphones in the array. The available data regarding optimal array and window sizes are limited and one of the objectives of this paper is to find appropriate parameters.

Defining $\mathbf{h}_i$ and $\mathbf{h}_j$ as the windowed RIRs of microphones i and j at an arbitrary time instant, the TDOA of an event $\tau_{i,j}$ between microphones i, j can be estimated by finding the delay that maximizes the cross-correlation $\mathbf{r}_{\mathbf{h}_i,\mathbf{h}_j}$

$$\tau_{i,j} = \arg\max \{\mathbf{r}_{\mathrm{h_i,h_j}}\}. \tag{1}$$

Assuming a sound field model in which only one broadband sound event arrives within the windowed responses, $\tau_{i,j}$ can be related to the geometrical properties of the array and direction of propagation of the sound event.

$$\tau_{i,j} = (\mathbf{m}_i - \mathbf{m}_j)^{\mathrm{T}} \frac{\mathbf{d}_p}{c}, \tag{2}$$

where $\mathbf{m}$ [3 × 1] refers to the position of the microphones in cartesian coordinates, $\mathbf{d}_p$ [3 × 1] refers to the direction of propagation of a single event in the windowed response, T denotes the transpose operation, and c refers to the speed of sound. This operation is repeated for each of the $N_{mp} = \frac{M(M-1)}{2}$ microphone pairs.

The time differences for each pair and the difference vectors for their positions are collected into the vector $\boldsymbol{\tau}$ [N_{mp} × 1] and matrix $\mathbf{V}$ [3 × N_{mp}], respectively. Then Eq. (2) can be rewritten as

$$\boldsymbol{\tau} = \mathbf{V}^{\mathrm{T}} \frac{\mathbf{d}_p}{c}. \tag{3}$$

By calculating the Moore-Penrose pseudoinverse $(\cdot)^+$ of $\mathbf{V}^{\mathrm{T}}$ the least-squares solution is obtained, resolving the direction of propagation of the event.

$$\mathbf{d}_p = (\mathbf{V}^{\mathrm{T}})^+ \boldsymbol{\tau} c. \tag{4}$$

Finally, the DOA vector $\mathbf{d}$ [3 × 1] is the opposite vector of the direction of propagation

$$\mathbf{d} = -\mathbf{d}_p. \tag{5}$$

The previous process is repeated for each sample of the measured RIRs, resulting in a matrix $\mathbf{D}$ containing the DOA for each sample. The reader is referred to [1] for further details regarding the algorithm. Throughout this work we used the implementation provided in the SDM Toolbox [18] for Matlab to conduct the presented investigations.

1.2 Pseudo-Intensity Vectors (PIV) Method

The DOA estimation can also be done using alternative approaches, provided that one DOA is assigned to each sample in the RIR. The Spatial Impulse Response Rendering (SIRR) method [19, 22] introduced the use of pseudo-intensity vectors for the estimation of narrow band directional information from B-format (First Order Ambisonics—FOA) RIRs. As opposed to SDM, SIRR further aims at dividing the RIR into a directional and diffuse component. More recently a higher order variant (HO-SIRR) was introduced [23], introducing the capability of identifying the direction of arrival of multiple events arriving simultaneously.

The original conception of SDM [1] only contemplated DOA analysis using the TDOA method. However, in the SDM Toolbox [18], Tervo et al. included the PIV analysis approach to generate DOA estimates to be used with SDM. The method is largely based on that used for the characterization of the directional sound field component in SIRR. However the SDM Toolbox only included analysis of broadband responses. While PIV analysis using multiple bands could be used in conjunction with SDM, to the best knowledge of the authors this has not been evaluated in the past. Additionally, in spite of the growing popularity of this analysis approach with SDM, only a few recent studies have analyzed its objective and perceptual performance [20, 24, 25], and the topic warrants further attention.

As with the TDOA method, the goal is to obtain one directional estimate for each sample in the RIR.

$$\mathbf{D}(n) = \begin{bmatrix} \hat{x}(n) \\ \hat{y}(n) \\ \hat{z}(n) \end{bmatrix} = h_w(n) \begin{bmatrix} h_x(n) \\ h_y(n) \\ h_z(n) \end{bmatrix} * \mathbf{w}(k) \tag{6}$$

where h_w(n) is the omnidirectional channel (W) of the B-format signal, which approximates the pressure RIR. The three components of the pseudo-intensity vectors are represented by h_x, h_y, and h_z and correspond to the figure-of-eight virtual microphones of the B-format signal aligned with the X, Y, and Z axes, respectively. The DOA estimates are convolved with a Hanning window**w**for smoothing. This convolution is effectively a low-pass filter on the DOA data, and the optimal size of the window is currently unknown.

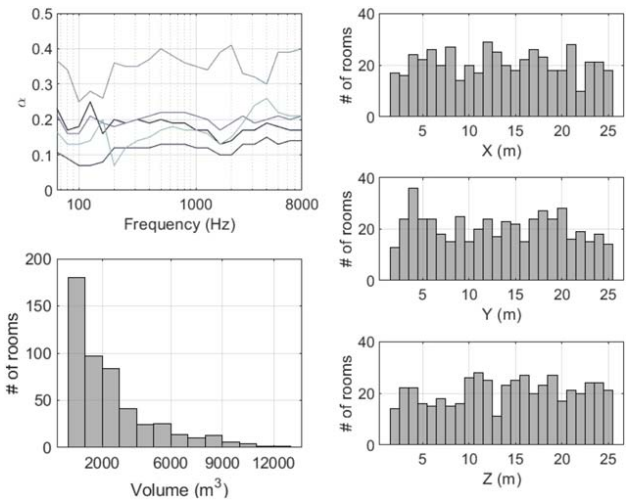


Fig. 1. Absorption data (top left), volume histogram (bottom left), and side-length histograms of the simulated rooms (right).

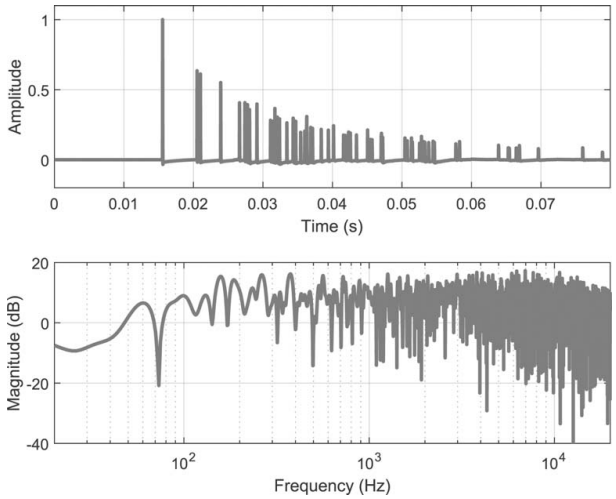


Fig. 2. ISM simulated RIR (top) and magnitude spectrum (bottom) for a room of dimensions $8 \times 6 \times 9$ m.

2 SIMULATIONS

In order to investigate the performance of the two presented approaches in idealized conditions, we simulated multichannel RIRs of 500 shoebox rooms. The wall lengths were randomly chosen using a uniform distribution containing lengths between 2 and 25 m, in order to cover a meaningful range of room sizes. The source and receiver locations were randomized as well in each simulation. Image Source Method (ISM) [26] simulations including frequency-dependent material absorption and air absorption were conducted using AKtools [27]. The materials for each wall are different and kept constant for all the room configurations. Room size distributions and absorption properties are shown in Fig. 1.

In order to allow for the evaluation of SDM with B-format signals, we expanded the functionality of the simulator to include ideal first-order microphones. The simulated RIRs contain 64 sound events, corresponding to the direct sound and specular reflections up to third order. Although the analysis we present in this section is not generalizable to the entire RIR, we decided to focus only on strong specular events for two reasons: it is known that the spatial analysis performed by SDM does not provide accurate results when multiple sound events start overlapping, as is the case in the late reverb [1], and the directionality of the late reverb is of limited perceptual relevance in common rooms [28]. An exemplary simulated RIR is presented in Fig. 2.

2.1 Evaluation Metric

We propose an objective metric ε_{DOA} to evaluate the performance of the DOA estimations. For each of the samples in the RIR we compute the angular distance between the ground truth DOA $\mathbf{D}_{ISM}(n)$ and estimated direction $\mathbf{D}(n)$. These are then weighted by the energy of each sample and normalized by the total energy of the RIR.

$$\epsilon_{DOA} = \frac{\sum_{n=1}^{N} \arccos\{\mathbf{D}(n)^{\mathrm{T}}\mathbf{D}_{ISM}(n)\}p(n)^2}{\sum_{n=1}^{N} p(n)^2} \tag{7}$$

where p_n represents the instantaneous pressure of sample n and N is the number of samples in the RIR. Note that the DOA vectors contained in the matrices $\mathbf{D}_{ISM}$ and $\mathbf{D}$ are expressed in cartesian coordinates and normalized to define unit vectors.

It is worth noting that the proposed metric relies on comparing the estimated DOA of each sample in the RIR. Thus it is only suitable for the case in which sound events in the RIR are not overlapping and each sample in the RIR has only one associated DOA.

2.2 Time Differences of Arrival (TDOA) Evaluation

As discussed previously, the requirements for estimation with time difference of arrival consist of a compact microphone array with at least four microphones arranged in a 3D space. If the data are intended for auralization, one of the microphones must be omnidirectional—for analysis only, multiple directivities are acceptable. These somewhat relaxed requirements resulted in a variety of experimental works using various array configurations, including microphone arrays arranged in orthogonal directions of various sizes—with or without a center microphone [7, 5, 9, 20, 6], a tetrahedral array with a physical or virtual omnidirectional microphone [15, 20], or a 12-element star-shaped array [20].

2.2.1 Array Size

Arrays composed of 6 or 7 omnidirectional microphones (3 orthogonal pairs, with or without a center microphone) seem to be among the most commonly used topologies [7, 20, 6, 4, 9]. However, to the best of our knowledge, no formal comparison between array topologies and dimensions has been completed to date.

Given the extended use of this topology, we chose to investigate the optimal size of this geometry. To that end we performed a DOA analysis using the function `SDMpar` from the SDM Toolbox [18] on the 500 simulated ISM

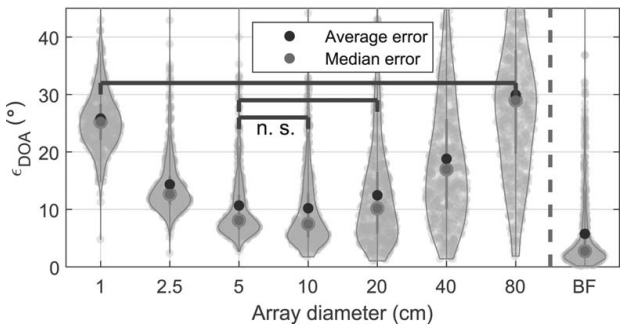


Fig. 3. DOA error as a function of microphone array diameter (500 ISM simulations at $f_s = 48$ kHz). The open array is a 7-microphone array with a center capsule and 6 capsules arranged as pairs in orthogonal directions. BF refers to ideal B-format signals. Brackets refer to statistically non-significant differences between groups ($p > 0.01$). The statistical analysis was conducted using a Kruskal-Wallis test with Tukey's range post-hoc correction (balanced dataset with non-normal distributions).

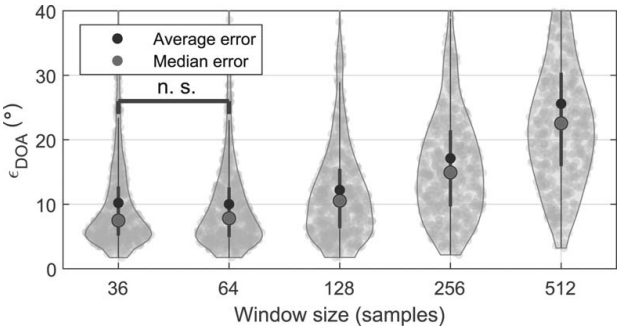


Fig. 4. DOA error as a function of window size (500 ISM simulations at $f_s = 48$ kHz for a 7-microphone open array of 10-cm diameter with a center capsule and 6 capsules arranged as pairs in orthogonal directions). Brackets refer to statistically non-significant differences between groups ($p > 0.01$). The statistical analysis was conducted using a Kruskal-Wallis test with Tukey's range post-hoc correction (balanced dataset with non-normal distributions).

RIRs. The results in Fig. 3 show that for a sampling rate of 48 kHz, an array with diameters of 5 and 10 cm results in the smallest DOA estimation errors among the evaluated dimensions. Smaller and larger sizes yield statistically significant increases in error.

The presented results partially confirm the findings of perceptual validations by Ahrens [20]. They found that at this sampling rate ($f_s = 48$ kHz), when comparing auralizations against a reference BRIR, arrays of 6 sensors with diameters of 10 and 20 cm result in smaller perceptual differences than arrays of 4-cm diameter or other configurations such as a tetrahedral array of 4.8-cm diameter or a 12-element array of 10-cm diameter.

It is worth highlighting the substantial difference between the average and median errors in all cases reported in Fig. 3. This suggests that the directional estimation error is especially high in some cases, leading to long tailed distributions.

When analyzing small spaces, such as a car cabin [7], it might be desirable to use smaller arrays to enable the use of smaller analysis windows. In these cases higher sampling rates might be necessary in order to avoid quantization in the resolved DOA estimates caused by insufficient time resolution. However a comparison between 48 and 96 kHz carried out using a compact tetrahedral array of approximately 4.8 cm of diameter (Core Sound TetraMic) found no significant benefit of increasing the sample rate when analyzing larger halls [15]. A formal comparison of array sizes at multiple sampling rates warrants more investigation. However the data presented here serve to lay a foundation and provide formal validation of the suitability of a relatively popular array topology used for SDM.

2.2.2 Window Size

The size of the sliding window used for DOA estimation theoretically governs the compromise between temporal and spatial resolution. While a larger window length would enable a more robust estimation of single events, it also increases the probability of multiple events arriving within the same window and consequently violating the soundfield model based on a succession of specular events. Tervo et al. [1] recommend the use of a window that is slightly longer than the time that it takes for one acoustic event to travel along the longest array dimension. In order to formally validate this recommendation we completed the DOA analysis and computed the estimation error using the 10-cm array configuration.

In Fig. 4 the DOA estimation errors [computed with Eq. (7)] for various window sizes are reported. As can be seen, shorter windows yield smaller errors that increase steadily with increasing window length. For the evaluated case of a 10-cm diameter array at 48 kHz, a window size of 36 samples seems most appropriate, although differences to the 64-sample window do not seem obvious. Thus we conclude that sizes between 36 and 64 samples are appropriate for this configuration. We hypothesize that fine tuning might provide a practical benefit depending on the structure of the specific analyzed RIR.

Theoretically, for the studied case, smaller windows could be used, as long as the window length is larger than the time needed for a plane wave to travel between the two most distant microphones. However the minimum value allowed in the SDM Toolbox is slightly higher and selected by default and thus for practical reasons we decided to limit the smallest size.

2.3 Pseudo-Intensity Vectors (PIV) Evaluation

The use of coincident array configurations (B-format arrays) has recently become more common, as it does not require specific open array topologies and the signals can be obtained in a variety of ways, either by using B-format arrays or taking subsets of Ambisonic signals in higher-order spherical arrays. However the rendering results generated with SDM and PIV DOA estimation have often been found to be unsatisfactory [24, 25, 23, 20]. For instance, in direct comparisons against a reference, SDM renderings from B-format arrays and PIV DOA estimation presented lower perceptual ratings than those from open array configura-

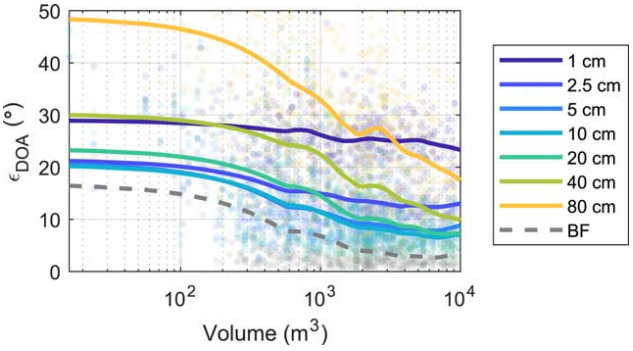


Fig. 5. DOA estimation error as a function of room volume for various array configurations. Circular markers represent individual observations; lines show a moving average (200 samples) of the observations.

tions and TDOA DOA estimation [20] or first-order SIRR renderings [23]. It is shown in [20] that the use of measured B-format RIRs results in poor directional estimates, even for samples of the direct sound. That could be partially attributed to the imperfect directional properties and spatial aliasing exhibited by B-format microphones above the spatial aliasing frequency. However it has also been suggested that the equalization process provided in the SDM Toolbox could be partially responsible for the generation of audible artifacts [24], as it suffers from time aliasing [20]. In addition, unlike with open microphone arrays, in the B-format estimation the DOAs are obtained from single-sample snapshots of the pseudo-intensity vectors and the estimates are prone to very fast variability. These effects can be partially mitigated by band-pass filtering and smoothing the DOA estimates, and some studies have reported that SDM auralizations from B-format arrays can indeed be perceptually very close to a reference [16]. Given the mixed results, we aim at evaluating the performance of broadband PIV estimation, as it is generally used in SDM.

Similar to our analysis with the open microphone arrays, we simulated 500 RIRs corresponding to shoebox rooms. In this case we defined ideal B-format microphones and used the function `SDMbf` (without windowing) to obtain the DOA estimates. As reported in Fig. 3, in an ideal simulation the results from a B-format array are approximately half an order of magnitude better than the best tested open array configuration. These results suggest that the B-format analysis might be preferable to TDOA. However in practice the quality of the results strongly depends on the A-to-B encoding and the quality of the reconstructed first-order directional patterns. This is further explored in Sec. 3.

2.4 Room Size

The DOA estimation error is presented as a function of room volume for different array sizes in Fig. 5. It can be seen that the estimation error tends to decrease with increasing room volume, with bigger arrays presenting more accentuated reductions in error. This is expected, as in larger rooms the average time between consecutive reflections is greater, thus reducing the probability of two reflections arriving simultaneously or within the same analysis window.

It is observed that the B-format array performs better than any open array counterparts at all room volumes. However, note that in this case the simulated B-format signals exhibit ideal directivities and are free from spatial aliasing, which is generally not the case in measured signals. For open arrays, very small and big arrays perform consistently worse than medium-sized arrays. This confirms our findings suggesting that medium-sized arrays (5, 10, and 20-cm diameter) are preferred at a sampling rate of $fs = 48$ kHz (see SEC. 2.2.1).

It is important to note once again that this analysis is based on limited-order ISM simulations without diffuse energy, thus representing the best case scenario for the analysis. We hypothesize that the same behavior would generalize to measured rooms, given that prominent reflections are already more spaced in time in larger spaces, but it is not possible to confidently generalize these findings from the presented results.

3 MEASUREMENTS

In addition to the aforementioned sound field assumptions, in simulations the array sensors exhibit ideal characteristics and the RIRs are free of noise. The PIV method resolves the DOA by providing an exact solution, while the TDOA method uses a pseudo-inverse, thus providing a least squares solution. We hypothesize that measurement noise, non-ideal microphone directivity limitations, and imperfections in the A-to-B format conversion might result in a noticeable analysis degradation, especially for the PIV method.

In order to compare the results of the TDOA and PIV approaches in a practical scenario, we conducted RIR measurements using a tetrahedral microphone (CoreSound Tetramic) in an apartment-like scene with a tall absorptive ceiling (see *FRL Apartment* in the *Replica* dataset [29]).

We used the A-format signals (four cardioid microphones at the vertices of the tetrahedron) to conduct the DOA analysis based on TDOA and the B-format signals for the PIV method. The A-to-B conversion is performed following the manufacturer's recommendations—using individually calibrated encoding matrices and the software *VVMic*. Note that the array is relatively small ($\sim$2 cm diameter) and thus not the optimal choice for the open array case. In addition the microphones are not omnidirectional but cardioid, potentially even further compromising the performance of the TDOA algorithm. However this enables a more accurate and direct comparison than repeated measurements with different arrays placed at the same position. Given that both analysis methods can be used with this array, a direct comparison aids in establishing guidelines for algorithm choice in practical scenarios.

Given the relatively complex scene geometry and large amount of furniture (see Fig. 6), reliable DOA ground truth data are not available. Thus we focus on a qualitative comparison of the results obtained using various window sizes.

Fig. 6. Top view of the room used for the measurements in the PIV and TDOA comparison. The visualization is part of the Replica Dataset [29] and the furniture was in a different configuration during the acoustic measurements. Approximate source (S) and receiver (R) locations are marked as black squares.

Fig. 7 contains a collection of spatial energy maps corresponding to the analyzed measurement. We focus the analysis on the first 20 ms of the RIR, as they contain several prominent reflections. Both methods present considerable agreement for the DOA of more energetic samples, with longer windows providing more stable estimates that result in energy clusters at discrete locations. A slight angular offset is apparent when comparing the two methods. Provided that the signals for the analysis come from the same measurement set, a possible explanation for this offset is a non-ideal encoding of the first-order directional patterns of the B-format array.

Given that the ground-truth data for the DOAs are not available, it is not straightforward to assess which of the estimations is closer to the actual DOAs. However it seems reasonable to conclude that raw DOA estimates for the PIV method (without a smoothing window) present significantly worse performance than when they are smoothed, with estimated DOAs of multiple highly energetic samples scattered around the entire sphere—including samples of the direct sound. This could result in noticeable localization artifacts if these data were to be used directly for auralization. Thus, in a practical scenario, a smoothing window should be used when using the PIV method. Note that in the PIV case the windowing acts as a low-pass filter on the pseudo-intensity vectors, as a Hanning window is convolved with the product of omnidirectional and velocity RIRs. Although the stability of the PIV analysis increases with longer windows, the DOA of the strongest events is not significantly affected by the window size.

For the TDOA approach, longer windows result in clustering many low-energy DOA samples into larger clusters, resulting in cleaner energy maps. However, note how when increasing the window size from 16 to 64 samples, the estimated DOA of one specific reflection changes drastically. Specifically, in the left column of Fig. 7, the reflection represented by the light green icons presents most of its energy around $[-160^\circ, 20^\circ]$ in the 16-sample plot, then moves to $[55^\circ, -40^\circ]$ in the 64-sample plot.

Another identifiable difference between the two methods is that while the overall stability improves for both methods with longer windows, in the PIV analysis the estimates follow continuous traces, creating trailing patterns between DOAs of strong events due to the effects of the window convolution. This is also somewhat present in TDOA estimates, although to a much smaller extent.

4 BINAURAL RENDERING

Binaural room impulse responses (BRIRs) can be synthesized as a weighted sum of HRTFs corresponding to each DOA, appropriately delayed and weighted by the amplitude of the instantaneous pressure of the omnidirectional RIR. We presented this method previously in [6]. Alternative implementations based on binaural Ambisonics and virtual loudspeaker layouts can be found in [16] and the SDM Toolbox [18], respectively.

The SDM sound field is defined by a [1 × N] vector $\mathbf{p} = [p_1, p_2, ..., p_N]$ containing the pressure RIR and a [3 × N] matrix $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_N]$ indicating the DOA for each of the samples in cartesian coordinates. The DOA matrix, $\mathbf{D}$, can be rotated to render BRIRs corresponding to arbitrary head orientations,

$$\mathbf{D}^u = \mathbf{R}_z(-\theta^u)\,\mathbf{R}_y(-\phi^u)\,\mathbf{D} \tag{8}$$

where $\mathbf{R_y}$ and $\mathbf{R_z}$ represent rotation matrices corresponding to the head orientation (θ^u, ϕ^u). Here a right-hand coordinate system is used with positive Y corresponding to left and positive Z to up.[1]

The indices $\hat{k}_n^u$ of the closest HRIRs for each sound event in each head orientation, u, are selected by finding the nearest HRIR for each sample, n, in the rotated DOA matrices $\mathbf{D}^u$.

$$\hat{k}_n^u = \underset{n\in 1,...,N}{\arg\min}\{d(\mathbf{D}_n^u, \hat{\mathbf{D}})\} \tag{9}$$

where $\hat{\mathbf{D}}$ is a [3 × K] matrix containing the source/receiver relative orientations of the HRIR dataset in cartesian coordinates and $d(\cdot, \cdot)$ is the Euclidean distance.

The BRIR for an arbitrary head orientation, $\mathbf{BRIR}^u$, is then constructed by delaying the HRIRs corresponding to indices $\hat{k}_n^u$ at the nth position by n samples and multiplying them by the instantaneous pressure p_n contained in the pressure RIR:

$$\mathbf{BRIR}^u(t) = \sum_{n=1}^{N} p_n\,\mathbf{HRIR}_{\hat{k}_n^u} \circledast \delta(t-n), \tag{10}$$

where $\mathbf{HRIR}$ is a three-dimensional [H × K × 2] matrix containing an HRIR dataset of H samples (per channel) and K source/receiver relative orientations. Samples in the BRIR are indicated by t.

To improve the timbral fidelity of the binaural reproduction, these rendered BRIRs can be further perceptually optimized by processing the DOA matrix $\mathbf{D}$ prior to the binaural rendering and performing reverberation equalization

[1]Note that the the DOA must be rotated in a reversed order to achieve a correct rotation. Roll rotation is excluded from the equation.

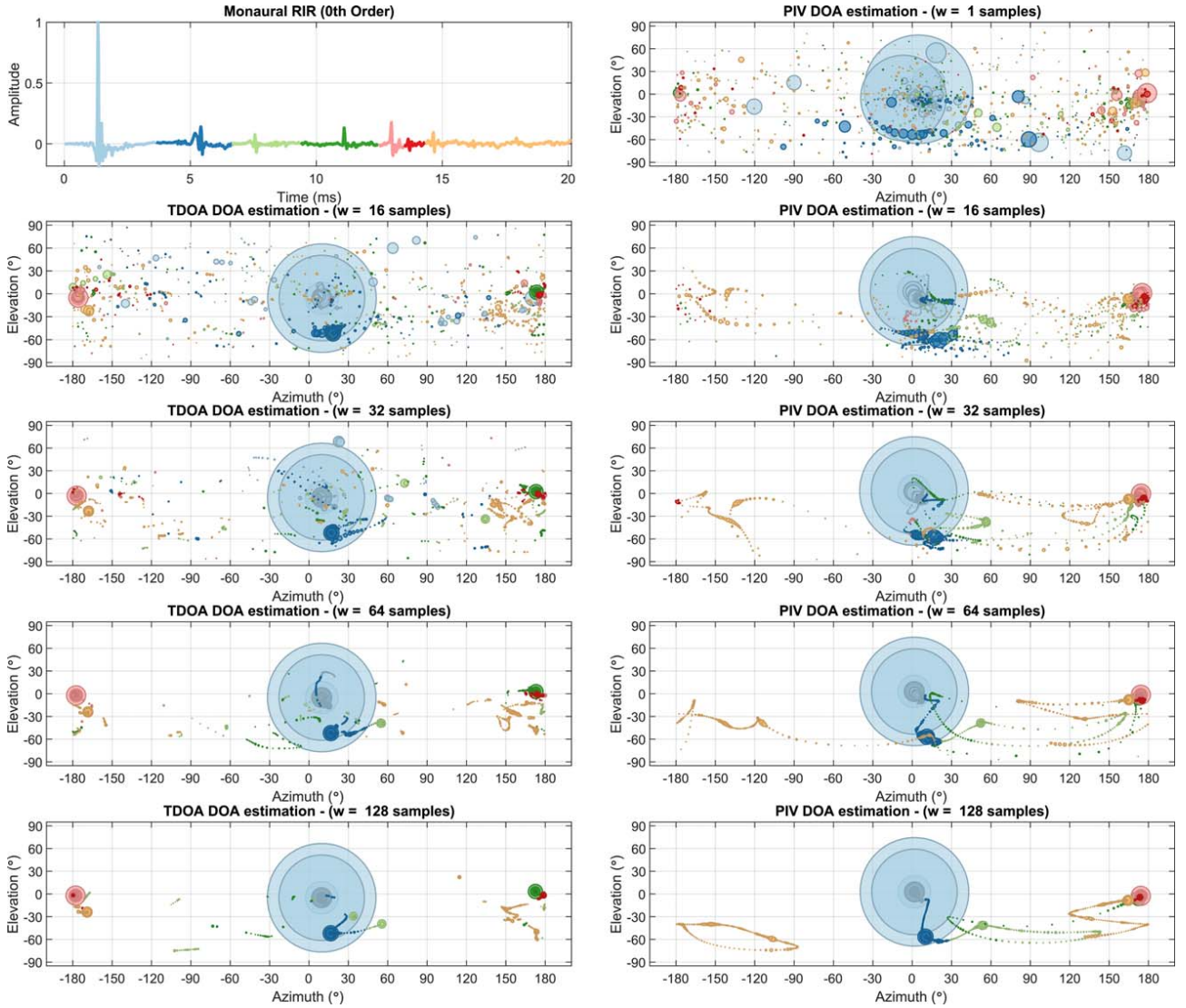


Fig. 7. Comparison of the DOA estimation results obtained from TDOA (left) and PIV (right) analysis at various window sizes ($f_s = 48$ kHz) using a tetrahedral array (Tetramic). Each circular marker represents a sample of the monaural RIR, with area proportional to the instantaneous energy of each sample. Note that in PIV windows are convolved with the signal, effectively low-passing the RIR.

on the rendered BRIRs. These processes are described in the subsequent sections.

4.1 DOA Postprocessing

In measured RIRs, sound events usually span multiple samples, as opposed to simulations, where events are usually very compact in time. As demonstrated in Sec. 3, when analyzing measured RIRs, it is common to obtain DOA estimates that fluctuate over the course of a single event. This can potentially result in spatial spread and spectral distortions of these events. Thus it is desirable to post-process the DOA estimates to minimize potentially audible artifacts.

In loudspeaker rendering, a common approach is the use of Nearest Loudspeaker Synthesis [7, 9], which assigns the DOA to the closest loudspeaker. While this reduces the spatial spread of single events by collapsing nearby DOA values to a single location, it might result in noticeable localization shifts, especially if the distance between the direction of the direct sound and the closest loudspeaker is larger than the minimum audible angle. An optimization method of the loudspeaker layout is available in [30].

When dealing with binaural synthesis, previous studies suggest that using a moving-average filter to smooth the DOA estimates is an effective post-processing approach [6, 16]. When using synthetic spatial data to auralize an omnidirectional RIR, it is desirable to use a certain degree of smoothing on the spatial data [21], resulting in smaller perceptual differences than when using random unfiltered data. Here we discuss potential alternatives to the post-processing of the DOA based on clustering of reflections and spatial quantization.

4.1.1 Direct Sound

As demonstrated in Fig. 7, the DOA of the most energetic samples in the RIR seem to be reliably estimated, although it is not uncommon to observe trailing patterns between those. Considering that in practice each acoustic event has a specific spectral shape, each event spans several

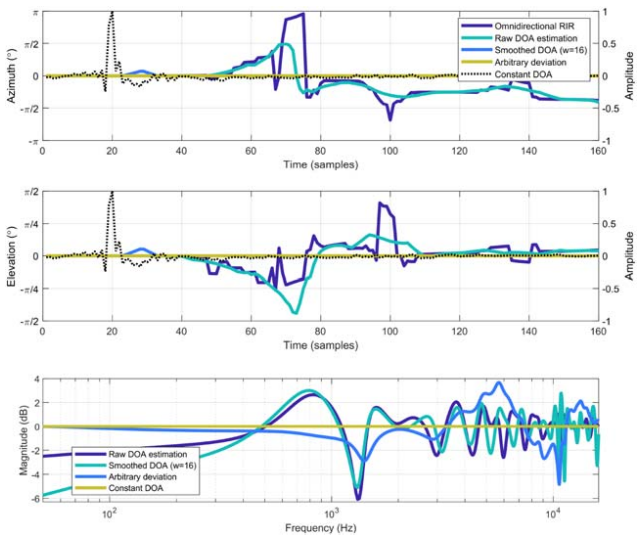


Fig. 8. DOA of the direct sound (top two panels) and spectral deviations resulting from mapping samples to multiple locations (bottom panel).

samples of the RIR. Thus mapping consecutive samples to disparate locations could potentially lead to spectral artifacts. To mitigate this and preserve the spectral properties of the direct sound we propose a post-processing step on the DOA matrix, enforcing a stable DOA for the direct sound.

We first locate the index ds of the sample with the largest amplitude in the RIR (direct sound)

$$ds = \arg\max\{|\mathbf{p}|\} \tag{11}$$

and then enforce the direction of the direct sound on the first i samples of the RIR

$$\mathbf{D}_p(n) = \begin{cases} \mathbf{D}(ds) & n \leq i \\ \mathbf{D}(n) & n > i \end{cases} \tag{12}$$

where $\mathbf{D}_p$ is the DOA matrix with constant DOAs for the first initial i samples and the original directions for the rest of the RIR.

Adjusting i in a case-by-case basis allows optimizing the trade-off between spectral and spatial fidelity. Theoretically the maximum allowed value of i is equal to the initial time delay gap (ITDG) of the RIR. Due to pre-ringing of each sound event, in practice i will be slightly lower. The goal is to maintain a constant DOA for the direct sound for as long as possible without distorting the DOA of the first reflection. Note that as low frequency components extend longer in time, in RIRs with longer ITDG the spectral distortion of the direct sound can be corrected to a greater extent.

To demonstrate the effect of this post-processing step in a practical scenario we auralized an RIR measured with a 7-microphone array (10-cm diameter, one central microphone) and compared the magnitude spectrum of the direct sound when auralized using the original DOA as estimated using the TDOA approach or when enforcing a stable DOA on first $i = 160$ samples.

The data are reported in Fig. 8, showing the DOA estimates in four different cases: raw DOA estimates as produced by the SDM algorithm (using minimum window size—36 samples), smoothed DOA estimates with a moving average filter of 16 samples, stable estimates with an arbitrary deviation for illustrative purposes, and a reference case with perfectly stable DOA. It is observed that even when the DOAs of the most energetic samples are appropriately estimated, spectral deviations of up to 6 dB are present. Note that in the example presented here the responses correspond to the left channel of a BRIR rendered with a very dense HRIR dataset (20,624 directions). The effects are likely dependent on the rendering configuration (loudspeaker layout or HRIR grid) and thus not easily generalizable. However it is expected that rendering setups with more spatial resolution will suffer from higher spectral distortions, as small fluctuations in the DOA result in samples being mapped at different locations.

4.1.2 DOA Quantization

The same spatial-timbral trade-off discussed for direct sound is present for early reflections and late reverberation. This becomes especially severe when low-passed and overlapped reflections appear in the RIR, rendering the high temporal resolution of the DOAs unusable. This is typically the case in common rooms, where air absorption and surface absorption tend to attenuate high frequencies more than low frequencies. In this case using all the available information results in spreading specular reflections onto multiple directions, leading to spatial and timbral degradations. We thus suggest a straightforward approach based on clustering of early reflections to reduce the spatial spread of early reflections.

There are a number of suitable methods for the spatial clustering:

- Virtual layout optimization as in [30]: The selected DOAs are chosen from weighted spatial energy maps derived using the original DOA data and pressure RIR. The advantage of this approach is that an adaptive grid allows for graceful downsampling of the DOA.
- Density Based Spatial Clustering (DBSCAN) [31]: This method can be used to identify portions of the RIR with meaningful DOA data and cluster them. The parts of the RIR in which there are no reliable DOA data can be rendered separately as diffuse components or arbitrary directions can be enforced to preserve spectral information. At the time of writing we obtained preliminary results related to a post-processing algorithm using DBSCAN—although these are out of the scope of this manuscript.
- Quantization using an arbitrary grid: Using sparse spatial grids for quantization is the equivalent of using finite fixed virtual loudspeaker layouts. Although they might not provide an optimal layout, the implementation is straightforward. Below we include the rendering steps for an arbitrary grid. We further ex-

plored the minimum grid resolution for a Lebedev grid in perceptual tests (see Sec. 5).

Defining $\mathbf{D}_Q$ as a matrix containing the directions of an arbitrary grid in cartesian coordinates, the original DOA matrix can be quantized by finding the closest directions in the quantized grid

$$q(n) = \underset{n\in i,\ldots,N}{\arg\min}\{d(\mathbf{D}(n), \mathbf{D}_Q)\} \tag{13}$$

where q refers to the indices corresponding to the closest directions. Then, following a similar approach to Eq. (12), a final matrix of quantized DOAs $\mathbf{D}_{pq}$ can be defined.

$$\mathbf{D}_{pq}(n) = \begin{cases} \mathbf{D}_{ds}(n) & n \le i \\ \mathbf{D}_Q(q(n)) & n > i \end{cases} \tag{14}$$

Note that in the matrix $\mathbf{D}_{pq}$ the direction of the direct sound is enforced to be constant, as described in Sec. 4.1.1. At this point, using $\mathbf{D}_{pq}$ as the input variable in Eq. (8) and solving Eqs. (9) and (10) results in a re-synthesized BRIR with the post-processed DOAs as explained in this section.

4.2 Reverb Equalization

Direct auralization of an RIR using DOA data to either map the energy to discrete loudspeakers or generate BRIRs [as in Eq. (10)] results in a perceivable spectral whitening of those parts of the RIR with unreliable DOA estimation. When DOA estimates fluctuate randomly, single-band limited sound events are mapped onto disparate locations, resulting in broadband sound events. This is especially important in the late reverberation tail, resulting in an increase of the reverberation at high frequencies [5] or in environments with high echo density, such as small rooms or a car cabin [7]. An analysis of this rendering artifact and a time-frequency equalization to compensate for it were introduced in [7]. This equalization approach uses the pressure RIR p as a reference to generate a time-varying filter for each of the rendered directions. This is especially useful when SDM is used for loudspeaker-based auralization, as only a relatively low number of directional streams need to be equalized. However in binaural rendering with dense HRTF datasets this approach becomes impractical from a computing and memory perspective.

In this section we introduce the RTMod and RTMod+AP methods, which correct the reverberation time by acting on the BRIRs directly without using directional feeds as an intermediate step. The main idea of RTMod is to decompose the BRIR into fractional octave bands, modify the energy envelope of each subband separately, and finally reconstruct the broadband BRIR. The RTMod+AP variant is based on the same concept, but it processes the output signals through a cascade of 3 Schroeder Allpass filters to increase the echo density of the late reverberation.

Note that this approach is specifically designed for rendering directly into binaural signals. When using Ambisonics as an intermediate format, time-frequency equalization can be done in the spherical harmonics domain [16].

4.2.1 RTMod Equalization

To generate the band limited components of the BRIR we use the same implementation of a perfect reconstruction filter bank [32] found in the SDM Toolbox [18]. Assuming that the time-frequency deviations of the rendered BRIRs with regard to the original pressure RIR are not time dependent we can manipulate the energy envelope of each subband by using exponential functions.

$$\mathbf{BRIR}^u_{\text{corr}}(t) = \sum_{f=1}^{F} \mathbf{BRIR}_{\text{corr}^u,f}(t) \tag{15}$$

$$\mathbf{BRIR}^u_{\text{corr},f}(t) = \mathbf{BRIR}^u_f(t) e^{-t(d_{1,f}-d_{0,f})} \tag{16}$$

where $\mathbf{BRIR}_{\text{corr}^u}$ is the corrected BRIR and f refers to each frequency band. The constants $d_{1,f}$ and $d_{0,f}$ determine the amount of correction of each subband envelope and are determined using the RT_{60} of the pressure RIR and the BRIR.

$$d_{0,f} = \frac{\ln(10^6)}{2\,RT_{60,\text{resynth},f}} \tag{17}$$

$$d_{1,f} = \frac{\ln(10^6)}{2\,RT_{60,\text{orig},f}} \tag{18}$$

where $\text{RT}_{60,\text{ resynth},f}$ and $\text{RT}_{60,\text{ orig},f}$ refer to the reverberation time of band f of the uncorrected BRIR and pressure RIR, respectively.

After applying RTMod equalization the reverberation time of the resulting BRIRs are within one JND unit, which is defined as 5% of the RT_{60} according to ISO 3382, at most frequencies (see and Sec. 4.3 and [6] for a more detailed analysis of the equalization). However, due to the violations of the sound-field model, the late reverberation presents a more coarse fine envelope, likely due to consecutive events interfering constructively and destructively. Through informal listening we concluded that these artifacts are largely negligible when auralizing continuous signals but audible when rendering highly impulsive sounds.

4.2.2 RTMod+AP Equalization

One way to reduce the signal-dependent quality of the late reverb is by increasing the echo density of the late reverberation to achieve a more smooth decay. The goal is to break up strong specular reflections formed by constructive interference of multiple reflections due to incorrectly estimated DOA into multiple reflections.

Allpass (AP) filters have been extensively used in audio for decorrelation [33–35] or artificial reverberation [36–39] for their ability to act as impulse expanders. As such, when AP filters are designed as Schroeder Allpass sections [36, 37], they can be effectively used to increase the echo density of the late reverberation without affecting its spectral properties. This results in an RIR with a smoother time envelope. Additionally, if both left and right channels are processed with the same filters, the Inter-Aural Cross Correlation (IACC) is left unaffected. We propose the use of a cascade of 3 Schroeder Allpass filters (see Fig. 9) to process the late reverb of the broadband corrected BRIRs (RTMod+AP).

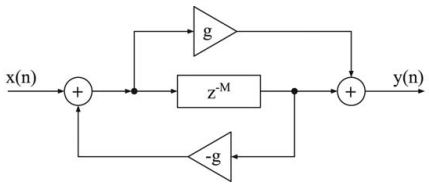


Fig. 9. Block diagram of a Schroeder Allpass filter.

The design of the Schroeder Allpass sections is based on two parameters—a delay (M) and gain (g). The values of these parameters can be largely based on artificial reverberation design. As such, delays must be coprime to minimize strong modulation effects. In the present paper we have obtained satisfactory results using delays of 37, 113, and 215 (on experiments run at a sampling rate of $f_s = 48$ kHz). The gain g can be set by using a desired reverberation time for the filters (RT_{60filt}).

$$g_{dB} = -60\frac{M}{f_s RT_{60filt}} \tag{19}$$

$$g = 10^{g_{dB}/20} \tag{20}$$

By setting relatively short reverberation times (in the order of 0.1 s) the RT60 of the BRIRs is largely unaffected by the Schroeder Allpass filters.

In order to process only the late reverberation, where spatial information is largely incorrect, we split the BRIRs at the mixing time [28] and the late reverb is processed using the Schroeder AP filter cascade. This effectively increases the diffuseness of the late reverberation without significantly changing its energy or IACC. Finally, the early reflections and processed late reverberation are summed back together using cosine ramps in the cross-fading region.

The choice of optimal number of filters and their parameters might be application dependent. Additionally the use of dynamic filter parameters could simplify the implementation and avoid explicitly dividing the BRIRs into early response and late reverberation. Adaptive filtering is used in SIRR to generate the time-varying diffuse component of the RIR. A similar application to the present approach warrants further research.

4.3 Instrumental Validation

To compare the performance of the equalization methods we compared a dummy-head reference measurement (KEMAR) with BRIR renderings generated using HRTF measurements of the same mannequin—source to the left (70° azimuth, 4° elevation). We used a microphone array of 10-cm diameter with a central microphone and 6 microphones arranged in pairs on orthogonal axes, an analysis window of 62 samples and a moving average window of 16 samples to smooth the DOA estimates. In the renderings we quantized the DOAs to 50 directions using a Lebedev grid while keeping the first 160 samples fixed to the original direct sound direction.

For the RTMod+AP equalization we used a mixing time of 3,800 samples (80 ms) and crossfade ramps of 1,024 samples. The filter delays were fixed to 37, 113, and 215 samples and their reverberation time was 0.1 s. The entire rendering process, from pressure RIR and DOA data to equalized RTMod+AP BRIRs, takes approximately 0.15 s on a laptop PC (Intel Core i7, 7th gen) running Matlab 2018 and Windows 10. In comparison the rendering using time-frequency equalization from [7] takes 8.9 s. Note that these refer to the rendering of one BRIR, corresponding to one arbitrary head orientation.

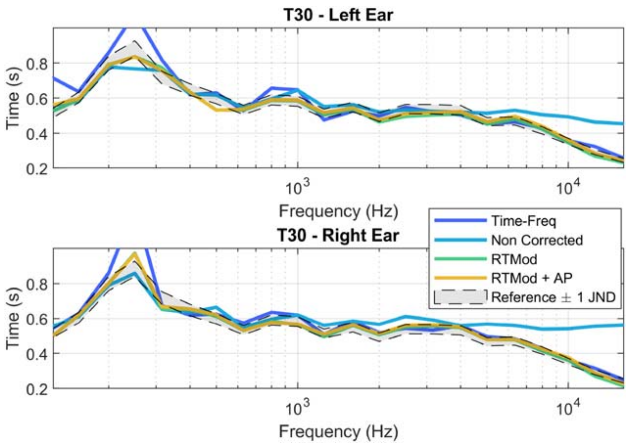


Fig. 10. Reverberation time (T30) of various reverberation equalization techniques.

Examples of estimated T_{30} for a BRIR processed with the presented methods are shown in Fig. 10. In this case, the reference T_{30} is obtained from the pressure RIR. It is clear that the non-corrected case, implemented as in Eq. (10), yields an excessive reverberation time above 4 kHz. The time-frequency equalization from [7] presents T30 results closer to the reference, although overestimated at low frequencies (approx. 250 Hz) and around 1 kHz. Finally, both RTMod [as in Eqs. (15) and (16)] and RTMod+AP methods present the closest results to the reference. Both RTMod and RTMod+AP present T_{30} errors smaller than the strictest accepted value of reverberation time JND (5% per ISO 3382-1:2009) over nearly the entire frequency range.

IACC has been linked to the perceived spatial quality of concert hall acoustics [40]. We computed the IACC for all the equalization methods on the full BRIR as well as early (0 to 80 ms) and late (80 ms to end) portions (see Fig. 11). Although discrimination thresholds and perceptual interpretation of IACC are topics of current research, we utilize a JND value of 0.075 as defined in ISO 3382-1:2009 for reference. The greatest deviations for all three methods are at low and mid frequencies, below 1 kHz. The time-frequency method presents the highest error at all ranges and portions of the RIR. Both RTMod and RTMod+AP methods provide a significant improvement, with deviations within ± 1 JND across the entire spectrum except at one band for the early part of the BRIR. At the late reverb deviations increase slightly at low frequencies. Note that the RTMod and RTMod+AP methods are almost equivalent in the early part of the BRIR, as the allpass filter cascade is only applied to the late reverberation (with a fade-in ramp). Additionally the negligible differences when comparing RTMod and

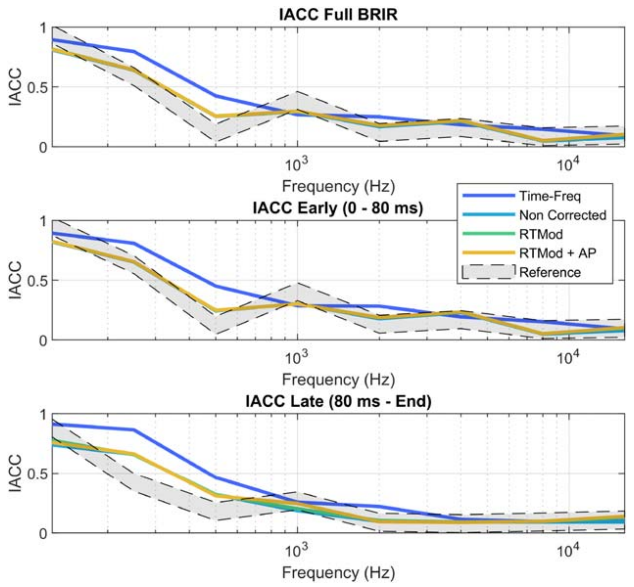


Fig. 11. Inter-Aural Cross Correlation of the reference and rendered BRIRs with various equalization methods.

RTMod+AP to the non-corrected BRIRs demonstrate that the spatial properties of the equalized BRIRs are preserved through the equalization method, which acts only on the time-energy properties.

Through this section we demonstrated the benefits of both RTMod and RTMod+AP over the uncorrected BRIRs and those with time-frequency equalization as in [7]. However neither T_{30} nor IACC analysis demonstrate the benefits of the AP addition to the RTMod equalization. As discussed extensively through the manuscript, a sound field description based on a succession of specular events is violated in the late reverberation. This results in a noisy time envelope, as the DOA estimates of the late reverberation are less reliable and multiple events interfere with each other.

Since allpass filters maintain the magnitude response of a signal while introducing changes in the phase, they can be effectively used to modify the fine time structure of a BRIR. In Fig. 12 we present the late reverberation amplitude envelopes for the left ear of the reference re-synthesized BRIRs. Although all the re-synthesized BRIRs present a noisier envelope than the reference, the use of allpass filters in the RTMod+AP method contributes to a smoothing of the envelope. In the studied case we used a cascade of three filters and we hypothesize that parameter tuning (number of filters, delays, decay time) could provide further gains. Informal listening revealed a significant improvement in the perceived similarity between the reference and RTMod+AP, as compared to the other methods. Refinements of the process and a formal perceptual evaluation are left for future work.

5 PERCEPTUAL EVALUATION OF SPATIAL QUANTIZATION

In previous sections we have objectively evaluated the effect of microphone array topology and proposed DOA post-processing and BRIR rendering alternatives to the original SDM implementation that allow rendering of BRIRs with high spatial density HRTF datasets. In this section we present a perceptual study in which we investigate the perceived plausibility of auralizations using DOA post-processing, as introduced in SEC 5.1 and RTMod equalization.

5.1 Implementation

In the experiment we conduct pairwise comparisons of real loudspeakers and renderings with various degrees of spatial resolution in the early reflections and late reverberation. The experiment was conducted in the same space used for objective comparisons (see Figs. 10–12 for acoustical parameters). We generated renderings using the RTMod equalization method (without AP cascade filtering), based on quantized DOA matrices using 7 grids of increasing resolution (1, 2, 6, 14, 26, and 50 points). The lowest resolution (1 point) collapses all the energy to the direction of direct sound. Grids with 2 and 6 points quantize the energy to left/right and left/right, top/bottom, and front/back, respectively. Larger grids (14, 26, and 60 points) are based on Lebedev grids. Spatial energy maps of each variant are shown in Fig. 13. The direct sound was fixed to the original direction for the first 128 samples in all the cases. The HRTF dataset used for rendering was from a KEMAR mannequin, obtained from boundary element method (BEM) simulations with 20,624 directions. The BRIRs were rendered with a resolution of 1° azimuth and 5° elevation.

The real-time rendering was done using a custom Max/MSP patch enabling dynamic rendering of 2 DOF (yaw and pitch) BRIRs. To save computational resources and memory, the reverberation was rendered separately and statically after a conservative mixing time (80 ms). Previous studies have shown that in typical rooms the dynamic rendering of late reverberation is not audible [28].

Tracking was implemented using an OptiTrack system with markers on both the listener and loudspeaker. This effectively enables pseudo-6DOF rendering, i.e., the relative angle between the loudspeaker and listener was always correctly tracked, thus rendering the correct direction for the direct sound. This ensures that small unintended translations of the subjects during the listening test do not result in perceivable localization shifts.

To enable direct comparisons between loudspeakers and binaural renders we used non-occluding headphones (AKG K1000). Although the occlusion from these headphones is arguably smaller than with generic on-ear or over-the-ear headphones, a comparison showed that they exhibit differences of 6 dB at 10 kHz when comparing HRTFs with and without headphones [41]. As the occlusion effect of the headphones is direction dependent, one possible way to compensate for them would be to render the BRIRs using HRIR datasets measured on subjects wearing the headphones. However as the stimuli were generated using generic HRTFs and including the headphone occlusion would not remove its effect from the real loudspeakers we decided not to include it. Informal listening revealed that

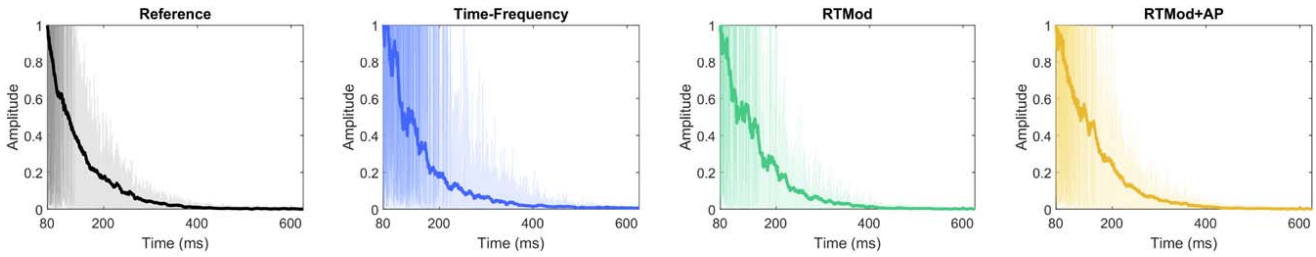


Fig. 12. Left ear absolute pressure signals (thin curves) and envelopes (thick curves) for the Reference BRIR (measured with a KEMAR dummy head) and various equalization methods applied to re-synthesized BRIRs.

the headphone occlusion did not affect the perceived location or spatial properties of the real loudspeakers and was only causing small coloration effects at high frequencies.

All the presented stimuli were band-passed between 200 Hz and 8 kHz using 15th-order Butterworth filters, which largely mitigated the effects of headphone occlusion at high frequencies. Generic headphone compensation filters based on KEMAR measurements using the same pair of headphones were used for all subjects. The filters were generated following the technique from [42].

5.2 Procedure

An increasingly relevant application of binaural rendering of room acoustics is the presentation of virtual sources in augmented reality scenarios along with real sources. Besides traditional similarity metrics, in the recent past dynamic binaural rendering has been evaluated in terms of plausibility [43] or authenticity [44]. While the definition of authenticity is unequivocal, i.e., the evaluated stimulus is perceptually indistinguishable from a reference under all listening conditions, plausibility experiments can be implemented in various degrees of strictness. For instance, in [43] and [45], the plausibility was assessed by using a yes/no test, in which listeners were asked to identify whether the stimulus was presented from a loudspeaker or binaurally using headphones. In this case listeners were hearing renderings corresponding to the room in which the experiments were carried out and were presented explicit versions of the real and virtual audio during the training phase. In comparison, in [46] the criterion is somewhat less strict, as listeners were presented with either simulations or measurement-based auralizations of real spaces and asked whether the stimuli corresponded to real or simulated rooms. In this case the listeners relied entirely on internal references related to plausibility of room acoustics and simulation artifacts that would differentiate real from simulated rooms. When utilizing only internal references, effects such as room acoustical divergence [47] or listener adaptation [48] could lead to changes in perceived externalization, thus affecting the plausibility ratings.

In order to account for plausibility at a stricter degree, we designed a 2-Alternative Forced Choice (2AFC) test in which in each trial subject was presented with two stimuli from the same location and asked to identify which of the two stimuli sounded more plausible. The concept of plausibility was discussed with the subjects and in this task it was equivalent to choosing which of the two simuli they thought corresponded to the sound generated by a real loudspeaker in the room (see Fig. 14). In each of the trials either both stimuli would be virtual or one of them would correspond to a real loudspeaker. This results in a test paradigm that combines plausibility based on the comparison to an internal reference (when two virtual sources are presented) or an explicit reference (when one of the sources is the real loudspeaker).

In order to eliminate potential influences due to visual elements or localization mismatch due to the use of generic HRTFs, the loudspeaker was hidden behind a curtain. The audio content used in the test was a sequence of castanets and only one single source location was used. Although the subjects were not given feedback or explicitly presented the real and virtual stimuli for comparison, they underwent an introduction to the experiment and conversation with the experimenter that allowed them to get acquainted with the natural acoustics of the room.

A group of eight expert listeners without known hearing problems participated in the test. All the subjects had previously participated in listening tests involving binaural audio and were familiar with room acoustic terminology. The total number of trials per subject was 21, resulting from all the possible pair combinations of 7 stimuli without repetitions, with 6 stimuli corresponding to re-synthesized BRIRs and 1 corresponding to the real loudspeaker in the room. The decision of not including repetitions responded to the fact that listeners were highly trained and we aimed at avoiding fatigue during the test.

To ensure subject reliability we provided listeners with unlimited time and instructed them to make use of natural head rotations in order to fully explore the sound scenes before making a decision. Listeners could switch back and forth between the two presented stimuli, which were played in sync and in loop. All the listeners heard the same stimuli, although the order of presentation was randomized. The collection of the responses was done using a touchscreen and GUI (see Fig. 14), minimizing the interaction between the subjects and experimenters.

5.3 Results

The results of the test are a decision matrix for each subject, corresponding to all the comparisons in the pairwise test design. By adding the rows of the matrix we can obtain the total number of selections of each stimulus. We use this

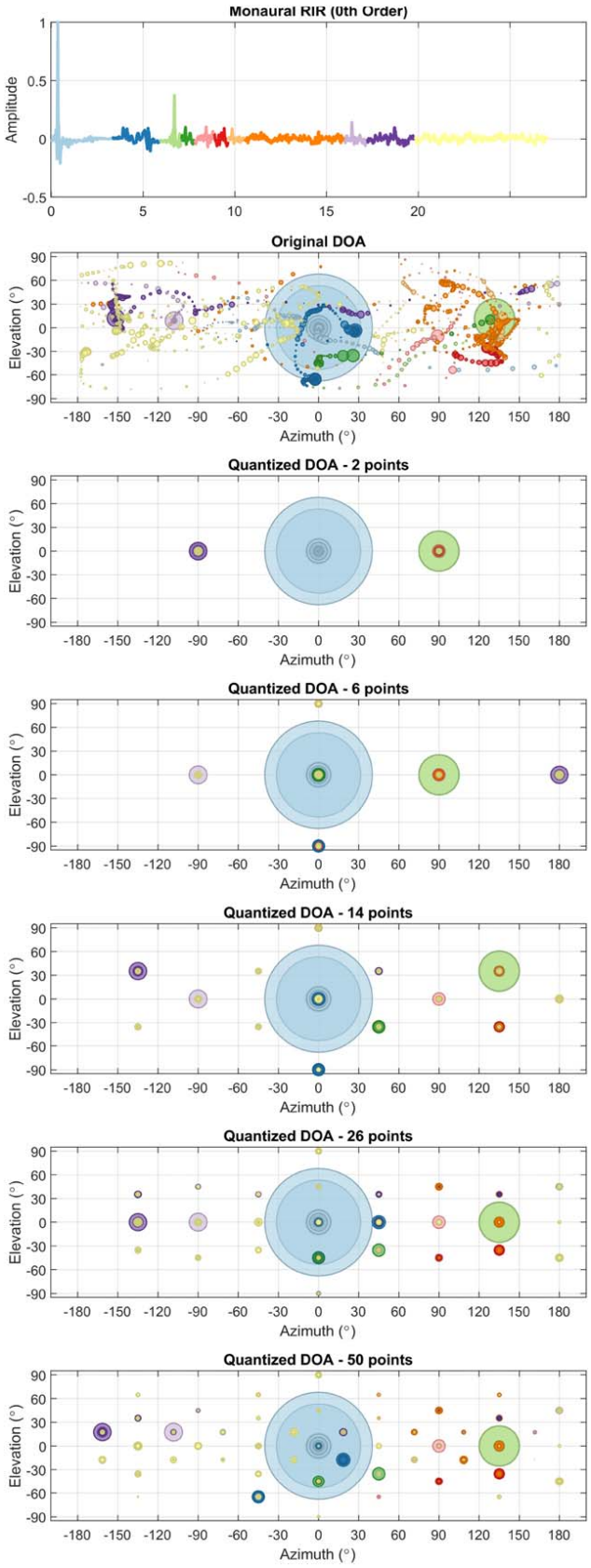


Fig. 13. Spatial energy maps of the renderings included in the listening test.

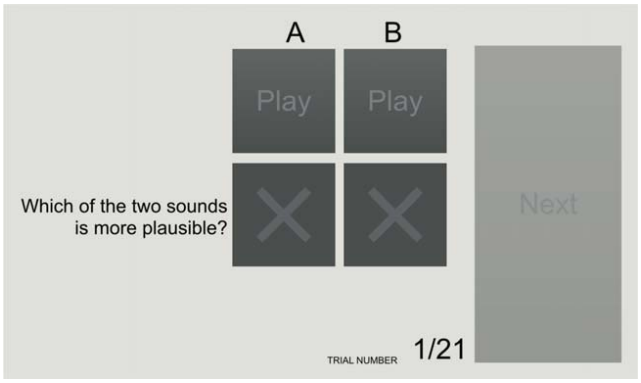


Fig. 14. GUI of the 2-AFC plausibility test.

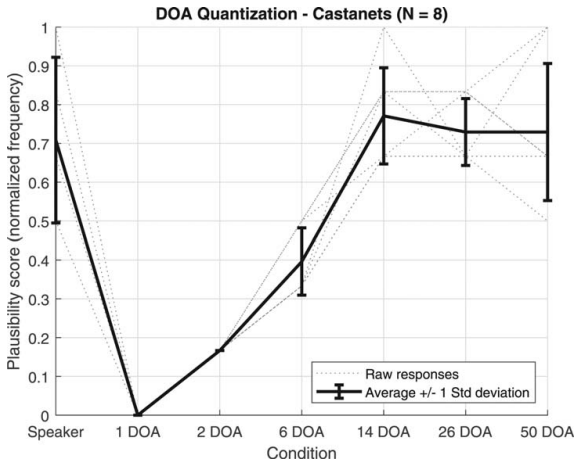


Fig. 15. Perceived plausibility of a real loudspeaker and SDM binaural renderings with various degrees of spatial resolution.

value and normalize it by the total number of presentations of each stimulus to obtain a 'Plausibility Score' P_i,

$$P_i = \frac{\sum_{j=1}^{N_s} a_{i,j}}{N_s - 1} \tag{21}$$

where $a_{i,j}$ is the response to comparison of stimuli $i$and j and has a value of 1 if i is selected as more plausible than j (and 0 if vice versa). N_s refer to the total number of stimuli (7 in this case). A P_i value of 1 would indicate that stimulus i is always selected as being more plausible than the rest in all cases. The results for each subject and a group average are shown in Fig. 15.

The results suggest that listeners perceive renderings with 14 or more DOAs as being as plausible as the real loudspeaker. The small spread at conditions 1, 2, and 6 DOA suggests that all listeners reliably discriminated between these conditions and the real loudspeaker or higher resolution renderings. In addition to showing that increased DOA resolution is not necessarily relevant in a practical scenario, the results suggest that the rendering improvements based on RTMod allow for the rendering of plausible virtual sources, even in the explicit comparison to real sources. Although the perceptual results in this specific scene are clear—there is no perceptual benefit in using more than 14 DOAs for rendering—it would be beneficial to conduct

similar tests in a variety of environments to ensure that they are generalizable.

6 DISCUSSION

The spatial decomposition method was initially developed for the research of concert hall acoustics [2, 3], and the public availability of a toolbox for Matlab caused a recent surge in its usage. It is nowadays applied in all kinds of room acoustics-related research, including car cabin acoustics [7, 8], stage acoustics [5, 14], audio-visual perception in virtual reality [12], speech intelligibility [11], dynamic binaural rendering [16, 6, 20], and acoustic preference in small rooms [9], among others. However it is a parametric method and despite its generalized use there is a lack of extensive perceptual validation and context-dependent optimization in the literature.

In an attempt to disentangle the effects of each step in the entire process and its application to binaural rendering we reviewed each stage separately. We focused on the case in which a sound field is generated in an enclosed space (small or medium room) by a single broadband source. The choice of microphone array and analysis parameters has a clear impact on the results of the DOA analysis. Although the B-format method can perform much better in simulations, the applicability to a real scenario might be case dependent and influenced by the encoding of the raw signals into B-format signals. Although we have not evaluated the possibility of performing a band-limited DOA estimation in several bands, this has been explored in the literature and applied in various studies [7, 9] using the TDOA approach. We want to note the fact that performing the PIV analysis in multiple bands would in fact converge to the same analysis procedure described in the (first order) SIRR method [19, 23] (although SIRR estimates a diffuse sound field component as well).

While using loudspeaker auralizations makes a comparison with a reference room difficult, binaural re-synthesis of an acoustic environment allows a direct comparison with dummy head recordings (or real spaces if non-occluding headphones are used). Recent studies have reported mixed results when using SDM auralizations. In various studies, authors presented satisfactory experimental results reporting perceptual ratings of SDM-based auralizations as being very similar to reference dummy head measurements [20, 16, 21, 17]. However all of these studies utilize custom implementations of the rendering part. In [20, 21], Ahrens explicitly mentions modifications to the time-frequency equalization to avoid perceivable time aliasing. In [16, 17], Zaunschirm et al. implement two variants, one based on a process similar to the one we describe in Eq. (10) and another based on Ambisonics upmixing. Studies using the original implementation found in the SDM Toolbox have reported more significant and case-dependent differences [24, 15, 23]. We thus want to draw attention to the equalization process as a critical factor in the quality of final renders.

We showed that both RTMod and RTMod+AP methods provide a substantial objective improvement as compared to the original time-frequency equalization. However they also increase the number of parameters in the rendering stage and may thus have a potential impact on the robustness of the method. The critical step is correctly estimating the reverberation time of both the original RIR and pre-corrected BRIRs. Although we have informally completed extensive perceptual validation of both variants of the proposed equalization we recognize the need for further formal evaluation including various acoustic spaces and content in order to evaluate the generalization of the observed improvements.

The fact that auralizations with spatial information quantized to 14 directions are perceived as equally plausible as a real loudspeaker suggests that the spatial resolution of the reverberation can be aggressively reduced without incurring perceptual degradations. Recent investigations led to similar conclusions when comparing Ambisonics renderings with full spatial resolution for the direct sound and reduced order for the reverberation [49]. A DOA clustering approach based on the perceptual relevance of prominent reflections could lead to further reduction of the needed number of directions used to render reflections.

7 CONCLUSIONS

In this paper we presented an optimization of SDM for the binaural auralization of multichannel RIRs, including the optimization of SDM analysis parameters, reduction of spatial resolution of the rendered BRIRs, and implementation of a new equalization approach for binaural renderings.

ISM simulations suggest that for a sampling rate of 48 kHz, an open array with a diameter close to 10 cm with an analysis window between 36 and 64 samples provides the lowest DOA error. For the case of B-format analysis, the results with simulations are significantly better than with an open array. However imperfections in A-to-B conversion in practical applications are not captured in simulations, and it is not possible to generalize the results from the B-format array to measurements.

Measurements with a Tetramic suggest that both TDOA and PIV methods are suitable to estimate the DOA of the strongest events in an RIR. When using PIV, longer convolution windows (lower low-pass cutoff frequencies) are preferred to obtain more stable DOA estimations. Choosing optimal window sizes might be case and array dependent and warrants more research.

We presented a reverberation equalization approach (RTMod+AP) composed of a Reverberation Time Modification (RTMod) step and an Allpass (AP) cascade filtering, yielding better objective results than the state of the art equalization for SDM at much lower computational cost.

Perceptual results suggest that equalization with RTMod provides perceptually plausible results when comparing dynamic binaural auralizations to real loudspeakers. Complete perceptual evaluation of RTMod+AP is left for future work.

The same perceptual experiments reveal that quantizing the early reflections and late reverberation DOA estimates to a Lebedev grid of 14 points does not result in perceptual

degradations when compared to denser grids, even with the use of a static late reverberation tail.

Future work includes the investigation of alternative methods to reduce the spatial resolution of DOA estimates and improve timbral preservation. Besides DOA quantization, time-varying or energy-informed clustering approaches could be explored to further reduce the spatial requirements without incurring perceptual impairments. A systematic analysis comparing both objective and perceptual performance of multiband directional analysis would help inform optimal parameters in a wider range of scenes. Finally full listening tests comparing the performance of RTMod+AP with other equalization approaches are also part of future work.

8 ACKNOWLEDGMENT

We want to thank Henrik Hassager, Nils Meyer-Kahlen, and Prof. Tapio Lokki for fruitful discussions on the work and valuable feedback on the manuscript. We are also indebted to the anonymous reviewers, whose contribution greatly improved the general quality of the paper.

9 REFERENCES

[1] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial Decomposition Method for Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 17–28 (2013 Jan.).

[2] 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, no. 2, pp. 842–857 (2013 Jan.), https://doi.org/10.1121/1.4770260.

[3] T. Lokki, J. Pätynen, A. Kuusinen, and S. Tervo, "Concert Hall Acoustics: Repertoire, Listening Position, and Individual Taste of the Listeners Influence the Qualitative Attributes and Preferences," *J. Acoust. Soc. Am.*, vol. 140, no. 1, pp. 551–562 (2016 Jul.), https://doi.org/10.1121/1.4958686.

[4] S. V. Amengual Garí, J. Pätynen, and T. Lokki, "Physical and Perceptual Comparison of Real and Focused Sound Sources in a Concert Hall," *J. Audio Eng. Soc.*, vol. 64, no. 12, pp. 1014–1025 (2016 Dec.), https://doi.org/10.17743/jaes.2016.0035.

[5] S. V. Amengual Garì, D. Eddy, M. Kob, and T. Lokki, "Real-Time Auralization of Room Acoustics for the Study of Live Music Performance," presented at the *Fortschritte der Akustik - DAGA 2016* (2016 March).

[6] S. V. Amengual Garí, W. O. Brimijoin, H. G. Hassager, and P. W. Robinson, "Flexible Binaural Resynthesis of Room Impulse Responses for Augmented Reality Research," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, pp. 161–166 (2019 Sept.), https://doi.org/10.25836/sasp.2019.31.

[7] 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, no. 11, pp. 914–925 (2015 Nov.), https://doi.org/10.17743/jaes.2015.0080.

[8] N. Kaplanis, S. Bech, S. Tervo, J. Pätynen, T. Lokki, T. van Waterschoot, and S. H. Jensen, "A Method for Perceptual Assessment of Automotive Audio Systems and Cabin Acoustics," presented at the *AES 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)* (2016 Jan.), conference paper 6-3.

[9] N. Kaplanis, S. Bech, T. Lokki, T. van Waterschoot, and S. Holdt Jensen, "Perception and Preference of Reverberation in Small Listening Rooms for Multi-Loudspeaker Reproduction," *J. Acoust. Soc. Am.*, vol. 146, no. 5, pp. 3562–3576 (2019 Nov.), https://doi.org/10.1121/1.5135582.

[10] S. Tervo, P. Laukkanen, J. Pätynen, and T. Lokki, "Preferences of Critical Listening Environments Among Sound Engineers," *J. Audio Eng. Soc.*, vol. 62, no. 5, pp. 300–314 (2014 May), https://doi.org/10.17743/jaes.2014.0022.

[11] O. Kokabi, F. Brinkmann, and S. Weinzierl, "Prediction of Speech Intelligibility Using Pseudo-Binaural Room Impulse Responses," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. EL329–EL333 (2019 Apr.), https://doi.org/10.1121/1.5099169.

[12] A. Saurì Suárez, N. Kaplanis, S. Serafin, and S. Bech, "In-Virtualis: A Study on the Impact of Congruent Virtual Reality Environments in Perceptual Audio Evaluation of Loudspeakers," presented at the *2019 AES International Conference on Immersive and Interactive Audio* (2019 Mar.), conference paper 67.

[13] J. Pätynen, S. Tervo, and T. Lokki, "Amplitude Panning Decreases Spectral Brightness With Concert Hall Auralizations," in *Proceedings of the AES 55th International Conference: Spatial Audio* (2014 Aug.), conference paper P-13.

[14] S. V. Amengual Gari, M. Kob, and T. Lokki, "Analysis of Trumpet Performance Adjustments Due to Room Acoustics," in *Proceedings of the International Symposium on Room Acoustics (ISRA)*, pp. 65–73 (2019 Sept.).

[15] S. V. Amengual Garì, W. Lachenmayr, and E. Mommertz, "Spatial Analysis and Auralization of Room Acoustics Using a Tetrahedral Microphone," *J. Acoust. Soc. Am.*, vol. 141, no. 4, pp. EL369–EL374 (2017 Apr.), https://doi.org/10.1121/1.4979851.

[16] M. Zaunschirm, M. Frank, and F. Zotter, "BRIR Synthesis Using First-Order Microphone Arrays," presented at the *144th Convention of the Audio Engineering Society* (2018 May), convention paper 9944.

[17] M. Zaunschirm, M. Frank, and F. Zotter, "Binaural Rendering With Measured Room Responses: First-Order Ambisonic Microphone vs. Dummy Head," *Appl. Sci.*, vol. 10, no. 5 (2020 Feb.), https://doi.org/10.3390/app10051631.

[18] S. Tervo and J. Pätynen, "SDM Toolbox for Matlab," https://www.mathworks.com/matlabcentral/fileexchange/56663-sdm-toolbox, accessed: 2020-06-29.

[19] J. Merimaa and V. Pulkki, "Spatial Impulse Response Rendering I: Analysis and Synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127 (2005 Dec).

[20] J. Ahrens, "Perceptual Evaluation of Binaural Auralization of Data Obtained From the Spatial Decomposition Method," in *Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 65–69 (2019 Oct.), https://doi.org/10.1109/WASPAA.2019.8937247.

[21] J. Ahrens, "Auralization of Omnidirectional Room Impulse Responses Based on the Spatial Decomposition Method and Synthetic Spatial Data," in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 146–150 (2019 May), https://doi.org/10.1109/ICASSP.2019.8683661.

[22] V. Pulkki and J. Merimaa, "Spatial Impulse Response Rendering II: Reproduction of Diffuse Sound and Listening Tests," *J. Audio Eng. Soc.*, vol. 54, no. 1/2, pp. 3–20 (2006 Feb.).

[23] 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, no. 5, pp. 338–354 (2020 May), https://doi.org/10.17743/jaes.2020.0026.

[24] C. Hold, *Spatial Decomposition Method on Non-Uniform Reproduction Layouts*, *Master's thesis*, TU Berlin (2019 Aug.).

[25] L. McCormack, A. Politis, O. Scheuregger, and V. Pulkki, "Higher-Order Processing of Spatial Impulse Responses," in *Proceedings of the 23rd International Congress on Acoustics*, pp. 4909–4916 (2019 Sept.).

[26] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950 (1979), https://doi.org/10.1121/1.382599.

[27] F. Brinkmann and S. Weinzierl, "AKtools—An Open Software Toolbox for Signal Acquisition, Processing, and Inspection in Acoustics," presented at the *142nd Convention of the Audio Engineering Society* (2017 May), convention paper 309.

[28] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual Evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 60, no. 11, pp. 887–898 (2012 Nov.).

[29] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, A. Clarkson, M. Yan, B. Budge, Y. Yan, X. Pan, J. Yon, Y. Zou, K. Leon, N. Carter, J. Briales, T. Gillingham, E. Mueggler, L. Pesqueira, M. Savva, D. Batra, H. M. Strasdat, R. De Nardi, M. Goesele, S. Lovegrove, and R. Newcombe, "The Replica Dataset: A Digital Replica of Indoor Spaces," *arXiv preprint arXiv:1906.05797* (2019).

[30] O. Puomio, J. Pätynen, and T. Lokki, "Optimization of Virtual Loudspeakers for Spatial Room Acoustics Reproduction With Headphones," *Appl. Sci.*, vol. 7, no. 12, p. 1282 (2017 Dec.), https://doi.org/10.3390/app7121282.

[31] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases With Noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, vol. 34, pp. 226–231 (1996).

[32] J. Antoni, "Orthogonal-Like Fractional-Octave-Band Filters," *J. Acoust. Soc. Am.*, vol. 127, no. 2, pp. 884–895 (2010 Feb.), https://doi.org/10.1121/1.3273888.

[33] G. S. Kendall, "The Decorrelation of Audio Signals and Its Impact on Spatial Imagery," *Comp. Music J.*, vol. 19, no. 4, pp. 71–87 (1995).

[34] M. Bouéri and C. Kyriakakis, "Audio Signal Decorrelation Based on a Critical Band Approach," presented at the *117th Convention of the Audio Engineering Society* (2004 Oct.), convention paper 6291.

[35] E. Kermit-Canfield and J. Abel, "Signal Decorrelation Using Perceptually Informed Allpass Filters," in *Proceedings of the 19th International Conference on Digital Audio Effects*, pp. 225–231 (2016 Sept.).

[36] M. R. Schroeder and B. F. Logan, "'Colorless' Artificial Reverberation," *J. Audio Eng. Soc.*, vol. 9, no. 3, pp. 192–197 (1961 Jul.).

[37] M. R. Schroeder, "Natural Sounding Artificial Reverberation," *J. Audio Eng. Soc.*, vol. 10, no. 3, pp. 219–223 (1962 Jul.).

[38] V. Välimäki, J. Parker, and J. S. Abel, "Parametric Spring Reverberation Effect," *J. Audio Eng. Soc.*, vol. 58, no. 7/8, pp. 547–562 (2010 Jul.).

[39] L. Dahl and J. -M. Jot, "A Reverberator Based on Absorbent All-Pass Filters," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)-00* (2000 Dec.).

[40] T. Okano, L. L. Beranek, and T. Hidaka, "Relations Among Interaural Cross-Correlation Coefficient (IACCE), Lateral Fraction (LFE), and Apparent Source Width (ASW) in Concert Halls," *J. Acoust. Soc. Am.*, vol. 104, no. 1, pp. 255–265 (1998 Jun.), https://doi.org/10.1121/1.423955.

[41] C. Pörschmann, J. M. Arend, and R. Gillioz, "How Wearing Headgear Affects Measured Head-Related Transfer Functions," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, pp. 49–54 (2019 Sept.), https://doi.org/10.25836/sasp.2019.27.

[42] J. G. Bolaños, A. Mäkivirta, and V. Pulkki, "Automatic Regularization Parameter for Headphone Transfer Function Inversion," *J. Audio Eng. Soc.*, vol. 64, no. 10, pp. 752–761 (2016 Oct.), http://doi.org/10.17743/jaes.2016.0030.

[43] A. Lindau and S. Weinzierl, "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acust. United Acust.*, vol. 98, no. 5, pp. 804–810 (2012 Sept./Oct.), https://doi.org/10.3813/AAA.918562.

[44] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the Authenticity of Individual Dynamic Binaural Synthesis," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 1784–1795 (2017 Oct.), https://doi.org/10.1121/1.5005606.

[45] C. Pike, F. Melchior, and T. Tew, "Assessing the Plausibility of Non-Individualised Dynamic Binaural Synthesis in a Small Room," presented at the *AES 55th Interna-*

tional Conference: Spatial Audio (2014 Aug.), conference paper 6-1.

[46] 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, no. 4, pp. 2746–2760 (2019 Apr.), https://doi.org/10.1121/1.5096178.

[47] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A Summary on Acoustic Room Divergence and Its Effect on Externalization of Auditory Events," in *Proceedings of the Eighth International Conference on Quality of Multimedia Experience (QoMEX 2016)* (2016 Jun.), https://doi.org/10.1109/QoMEX.2016.7498973.

[48] F. Klein, S. Werner, and T. Mayenfels, "Influences of Training on Externalization of Binaural Synthesis in Situations of Room Divergence," *J. Audio Eng. Soc.*, vol. 65, no. 3, pp. 178–187 (2017 Mar.), https://doi.org/10.17743/jaes.2016.0072.

[49] I. Engel, C. Henry, S. V. Amengual Garí, P. W. Robinson, D. Poirier-Quinot, and L. Picinali, "Perceptual Comparison of Ambisonics-Based Reverberation Methods in Binaural Listening," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, pp. 121–126 (2019 Sept.), https://doi.org/10.25836/sasp.2019.11.

THE AUTHORS


Sebastià V. Amengual


Johannes M. Arend


Paul Calamia


Philip Robinson

Sebastià V. Amengual is currently a research scientist at Facebook Reality Labs Research working on room acoustics, spatial audio, and auditory perception. He received a Diploma Degree in Telecommunications with major in Sound and Image in 2014 from the Polytechnic University of Catalonia (UPC) in 2014, completing his Master's Thesis at the Norwegian University of Science and Technology (NTNU). His doctoral work at the Detmold University of Music focused on investigating the interaction of room acoustics and live music performance using virtual acoustic environments. His research interests lie in the intersection of audio, perception, and music.

•

Johannes M. Arend received a B.Eng. degree in Media Technology from HS Düsseldorf, Düsseldorf, Germany, in 2011 and an M.Sc. degree in Media Technology from TH Köln, Cologne, Germany, in 2014. Since 2015, he has been a Research Fellow and working toward a Ph.D. degree at TH Köln and TU Berlin in the field of spatial audio with a focus on binaural technology, auditory perception, and audio signal processing. Between September 2019 and March 2020 he was a research intern at Facebook Reality Labs Research.

•

Paul Calamia is a research scientist on the Audio Team at Facebook Reality Labs Research, where he conducts research in room acoustics for augmented-reality applications. Previously he was a member of the Technical Staff at MIT Lincoln Laboratory in the Bioengineering Systems and Technologies Group and the Advanced Undersea Systems and Technology Group. His other prior positions include Assistant Professor in the Graduate Program in Architectural Acoustics at Rensselaer Polytechnic Institute in Troy, NY, Consultant and Head of R&D at Kirkegaard Associates in Chicago, IL, and Acoustical Engineer at Wyle Laboratories in Arlington, VA. He holds a bachelor's degree in mathematics from Duke University, a Master's degree in electrical and computer engineering from the Engineering Acoustics Program at the University of Texas at Austin, and a Ph.D. in computer science from Princeton University.

•

Philip Robinson is a research science manager in room acoustics and auditory perception at Facebook Reality Labs Research (FRL Research) in Redmond, WA. Prior to joining FRL Research, he incorporated virtual acoustics simulation and reproduction systems into building design processes at the architecture firm of Foster + Partners. He was a Fulbright Scholar and post-doctoral researcher at Aalto University in Finland, where he studied perception of concert hall acoustics, spatial auditory resolution, and echo thresholds. He has been a visiting researcher at EPFL in Switzerland and Hanyang University in South Korea. He received a Ph.D. from Rensselaer Polytechnic Institute in Troy, NY in 2012. In a previous life, he was a registered architect in his home state of New Mexico. He remains passionate about architecture, the study of which gave him a great interest in perception of environments, real or virtual.

3.3 A Reactive Virtual Acoustic Environment for Interactive Immersive Audio

Arend, J. M., Lübeck, T., & Pörschmann, C. (2019). In *Proc. of the AES International Conference on Immersive and Interactive Audio (IIA), York, UK* (pp. 1–10).

Audio Engineering Society

Conference Paper 9

Presented at the Conference on
Immersive and Interactive Audio
2019 March 27 – 29, York, UK

A Reactive Virtual Acoustic Environment for Interactive Immersive Audio

Johannes M. Arend[1,2], Tim Lübeck[1], and Christoph Pörschmann[1]

[1] *Institute of Communications Engineering , TH Köln - University of Applied Sciences, D-50679 Cologne, Germany*
[2] *Audio Communication Group, Technical University Berlin, D-10587 Berlin, Germany*

Correspondence should be addressed to Johannes M. Arend (Johannes.Arend@th-koeln.de)

ABSTRACT

Reactive virtual acoustic environments (VAEs) that respond to any user-generated sound with an appropriate acoustic room response enable immersive audio applications with enhanced sonic interaction between the user and the VAE. This paper presents a reactive VAE that has two clear advantages in comparison to other systems introduced so far: it generally works with any type of sound source, and the dynamic directivity of the source is adequately considered in the binaural reproduction. The paper describes the implementation of the reactive VAE and completes the technical evaluation of the overall system focusing on the recently added software components. Regarding the use of the system in research, the study briefly discusses challenges of conducting psychoacoustic experiments with such a reactive VAE.

1 Introduction

Imagine you could explore the acoustics of a virtual room through hand clapping, speaking, or whistling. Or you could rehearse in a virtual concert hall. Such interactive immersive audio applications can be implemented with a *reactive* virtual acoustic environment (VAE), meaning a real-time system that responds to user-generated sound, or more generally to any acoustic excitation of the user, with an appropriate acoustic room response. Up to now, several reactive VAEs have been presented in literature, like for example various virtual performance spaces for musicians [1, 2] or systems that reproduce one's own voice in a VAE [3]. In research, reactive VAEs have been used for instance to investigate the influence of room acoustics on music performance [2], to study how blind humans use echoes of self-generated oral sounds to explore the space around them [4], or to examine the effect of self-generated sounds (e.g. voice) on presence [3].

In general, the various reactive VAEs work pretty much the same, although the field of application and the implementation differ. Simply put, such a system captures the user-generated sound with one or more microphones, feeds the microphone signal(s) back into the VAE, and reproduces the corresponding acoustic room response with a loudspeaker- or headphone-based spatial audio system. However, most reactive VAEs are solely designed for a specific use case, like for example the reproduction of a certain musical instrument or one's own voice. Moreover, to our knowledge, none of the systems introduced so far assumes that the sound

This work was funded by the German Federal Ministry of Education and Research (BMBF 03FH014IX5-NarDasS)

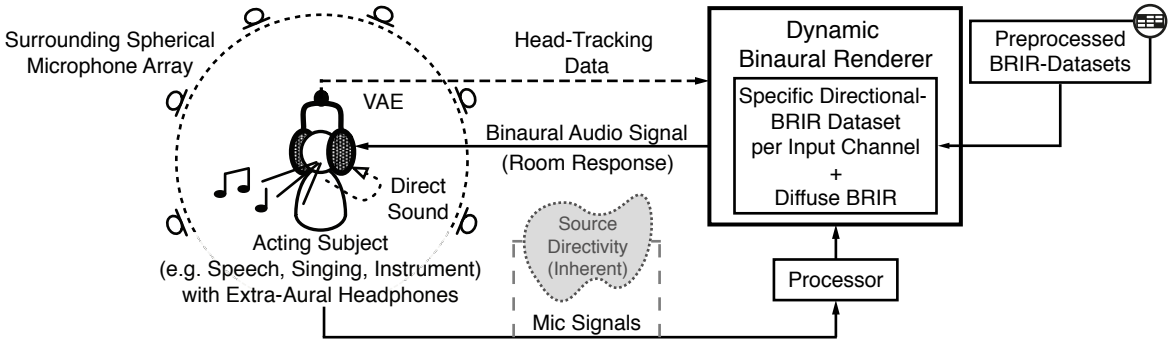


Fig. 1: Functional schematic of the reactive VAE including the main components of the system.

source (the user or the instrument) has a dynamic directivity. Thus, they neglect the specific changes in radiation dependent on phonemes (voice) or played notes (instruments) or due to movements of the user [5, 6]. From a technical point of view, neglecting the dynamic directivity leads to an inaccurate reproduction of the room response. Furthermore, several studies showed that auralizations (in a common VAE) with a dynamic directivity are perceived as more plausible and more realistic than auralizations with a static directivity [5, 6]. Thus, it is very worthwhile to incorporate the dynamic directivity in a reactive VAE.

In this paper, we present a reactive VAE that considers the dynamic directivity of the user-generated sound and, due to the system design, generally works with any type of sound source. The system is based on a surrounding spherical microphone array to capture the direction-dependent sound of the user, and on a dynamic binaural renderer for headphone-based reproduction of the corresponding room response. In Sec. 2 of this paper, we outline the basic idea of this reactive VAE. Sec. 3 describes the design and implementation of the system from the hardware and software side. In Sec. 4 we present a technical evaluation of the system focusing on the recently added software components. As the implementation was an ongoing process, further technical aspects have already been discussed in our previous publication [7]. Finally, Sec. 5 concludes the paper with a short summary and a brief discussion on challenges of conduction psychoacoustic experiments with the reactive VAE.

2 Basic Idea

The purpose of the system is to provide an adequate room-related reproduction of any user-generated sound in a headphone-based VAE. Consequently, the user, or in other words the acting subject, is central to the approach, as can be seen in the functional schematic illustrated in Fig. 1. To begin with, the user generates an arbitrary sound. A spherical microphone array surrounding the user captures this direction-dependent sound in real time (see Sec. 3.1). The resulting microphone signals, which inherently provide the dynamic source directivity, now go to a stand-alone software module simply called Processor (see Sec. 3.4). The module has several functions: On the one hand, it estimates the position of the sound source inside the array and levels each microphone signal dependent on the distance between the estimated source position and the respective microphone position. On the other hand, it additionally provides an adaptively filtered mono signal at the output, further called the reverberation source signal, which then is used as the excitation signal for the diffuse reverberation synthesis. The subsequent dynamic binaural renderer convolves each (leveled) microphone signal with a specific directional BRIR (BRIR - Binaural Room Impulse Response) and the reverberation source signal with a single diffuse BRIR. The directional BRIRs are preprocessed impulse responses describing a room-related direction-dependent binaural reflectogram per microphone channel (see Sec. 3.2). Since the VAE is based on dynamic binaural synthesis, the renderer requires an appropriate dataset of directional BRIRs per microphone channel. The diffuse BRIR describes the isotropic binaural reverberation of the simulated room (see Sec. 3.3). Finally, the resulting binaural audio signal, which is composed of direction-dependent reflections (without direct sound) and diffuse reverberation, is presented to the user over extra-aural headphones. By the use of such headphones,

the direct sound is maintained and reaches the ear of the user more or less unaffected, where it merges with the corresponding artificial room response. As usual for dynamic binaural synthesis, the renderer generates the binaural room response with respect to the head orientation of the user, which is provided by a head-tracking device.

3 Design and Implementation

The presented reactive VAE combines several newly developed hardware and software components with available standard components. The following sections describe the design and implementation of the main components specially developed for this application. Please note that, as the implementation was an ongoing process, several components are already described in greater detail in a previous publication [7].

3.1 Microphone Array

The surrounding spherical microphone array serves to capture the direction-dependent sound radiated by the user. The basic shape of the array is a pentakis dodecahedron with a diameter of 2 m. The 32 Rode NT5 microphones are located at the vertices of this shape on a constant radius of 1 m. The entire construction stands on stilts with a height of about 0.20 m, resulting in an array center height of about 1.20 m. To avoid reflections, the microphone holders (connectors) and the stilts are covered with foam absorbers. Fig. 2 shows a picture of the array, placed in the anechoic chamber at TH Köln, with a musician inside.

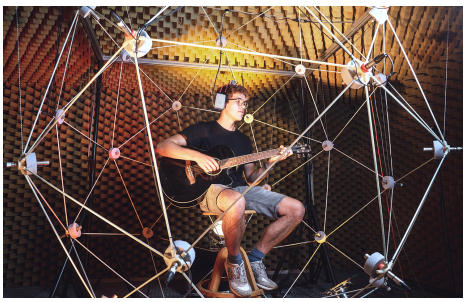


Fig. 2: Surrounding spherical microphone array used to capture the direction-dependent sound.

3.2 Synthesis of Directional BRIRs

Basically, each directional BRIR describes a direction-dependent room response by means of specular reflections. To obtain the required reflection pattern, the respective room is simulated with RAVEN [8] applying a combination of the image-source method and ray tracing. Depending on the application, the omnidirectional sound source and the receiver are placed at (almost) the same position (e.g. speech) or at slightly different positions (e.g. instrument). The results table of the image-source simulation provides all parameters necessary for the BRIR synthesis. Such a table contains the delay, the outgoing angle from the source, the angle of incidence at the receiver, and the frequency-dependent damping factors of each audible image source. All further processing including the actual synthesis is implemented in Matlab.

In a first step, each incident reflection is assigned to a specific direction of sound radiation. For this, the array geometry is segmented with respect to the microphones, resulting in a sphere with 32 faces (see Fig.3 (a)). As illustrated in Fig. 3 (b), the outgoing sound rays are now (notionally) surrounded with the segmented sphere, with the source placed at the center of the sphere. In a next step, every outgoing ray is assigned to a segment through intersection point calculation (see Fig. 3 (c)). Thus, the basic principle is to assign every outgoing sound ray, which later leads to an incident reflection, to one of the segments. The result is a list of the related incident reflection rays per segment (see Fig. 3 (d), the array geometry is only depicted here for a better understanding). In other words, to each microphone, these reflections are assigned that would occur when the room would be excited only through the respective segment with an ideal loudspeaker, directed towards the center of the segment and with a directivity according to the solid angle of the segment.

After the assignment, synthetic directional BRIRs per segment are generated by summing delayed, intensity-scaled, and filtered head-related impulse responses (HRIRs). The applied HRIRs were measured with a Neumann KU100 dummy head on a Lebedev grid with 2702 sampling points [9]. Because arbitrary directions are needed for the reflection synthesis, the HRIR dataset is stored as spherical harmonics coefficients, allowing to extract any required direction through spherical harmonics interpolation. To consider the absorption properties of the room, each HRIR is filtered with a specific reflection filter. The filters are based on the

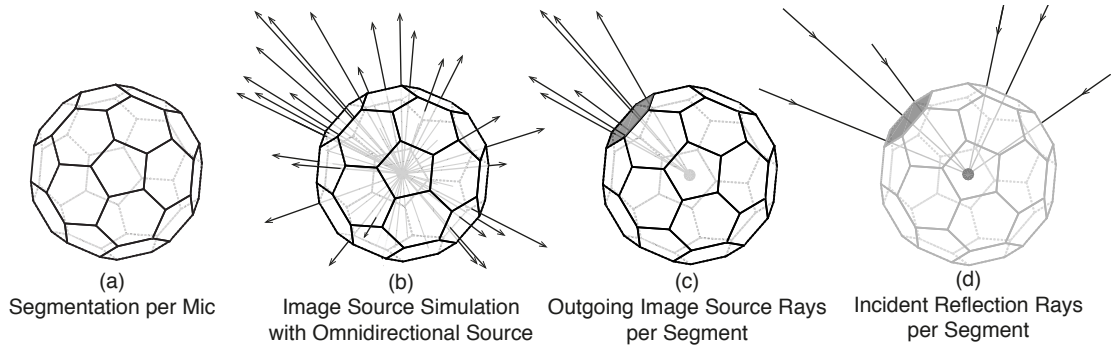


Fig. 3: Illustration of the ray-segment assignment process. The basic principle is to assign every outgoing sound ray, which later leads to an incident reflection, to a predefined segment of the microphone array.

frequency-dependent damping factors obtained with the simulation and are designed as minimum or linear phase FIR filter with adjustable length. Thus, the filter type and the filter kernel size can be chosen appropriately. Moreover, audio-latency compensation can be applied in this context, simply by subtracting the previously determined audio-latency value from the delay value of each reflection. For a receiver and source height of about 1.20 m, though, the first reflection is usually the floor reflection with a delay of about 7 ms. Depending on the buffer size, the audio latency is mostly higher than this, which is why optionally, the floor reflection can be skipped in order to provide full audio-latency compensation, or it can be maintained to compensate only the delay up to the first reflection.
The described synthesis procedure is repeated for each head orientation according to the applied spatial grid. Theoretically, each BRIR could be synthesized on a full-spherical grid. However, in most cases, a grid covering the horizontal plane (e.g. $0^\circ \leq \phi \leq 360^\circ$, steps of 2°) and parts of the median plane (e.g. $-30^\circ \leq \theta \leq 30^\circ$, steps of 10°) is entirely sufficient. Fig. 4 (top) again summarizes the entire processing chain.

3.3 Synthesis of Diffuse BRIR

As the term diffuse implies, we assume that the reverberant part of the virtual room is homogeneous and isotropic. Consequently, only one appropriate binaural reverberation impulse response needs to be applied. As indicated in Fig. 4 (bottom), the (raw) binaural reverberation can be acquired by dummy head measurements, by synthesis [10], or by simulation (using the result of the ray-tracing simulation). Depending on the input, the processing in Matlab involves different steps.

If a measured BRIR is used, the diffuse part after the (perceptual) mixing time is extracted and energy-matched to the simulated reverberation, which serves as the level reference. Then, the entire diffuse BRIR is shifted in time to the starting point of the first (latency-compensated) reflection. Finally, a linear or cosine-squared ramp is applied from the starting point to the perceptual mixing time. This way, the diffuseness already builds up through the early part, leading to significantly better perceptual results compared to when the early part only contains specular reflections [10]. Using fully synthetic reverberation is another option. In this case, the binaural reverberation is based on frequency-dependent shaped noise, matched in interaural coherence [10]. Generating the synthetic reverberation only requires the frequency-dependent reverberation time, provided by measurements or by the simulation. Similar to the processing of the measured BRIR, the synthetic reverberation is energy-matched, shifted in time, and faded in. The simplest way to acquire the diffuse BRIR though is to use the results of the ray-tracing simulation, since no further processing has to be applied.

3.4 Processor

As briefly described in Sec. 2, the reproduction system renders the directional and the diffuse part of the binaural room response separately. Ideally, when the sound source is exactly in the center of the array, the raw microphone signals can be passed straight to the binaural render for convolution with the respective directional BRIR. In practice, however, the user moves inside the array, which can lead to off-center shifts of the sound source. As a consequence, the reflections per segment are presented at incorrect levels (increased

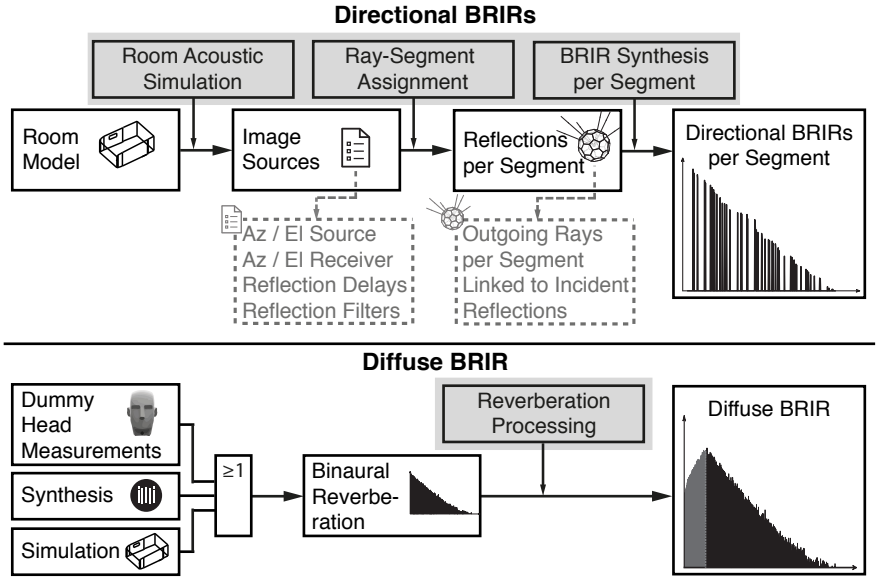


Fig. 4: Processing chain to synthesize the directional (top) and the diffuse (bottom) BRIRs.

or decreased according to the distance between microphone and sound source). To avoid this, the level of the microphone signals needs to be continuously adjusted according to the 1/r law, which again requires knowledge on the position of the sound source. For this purpose, we implemented two C++ software modules called Localizer and Leveler, which are part of the Processor.

As the names suggest, the Localizer determines the position of the sound source, and the Leveler adjusts the level of each microphone signal dependent on the distance between the estimated source position and the respective microphone position. In brief, the Localizer follows a TDOA-based approach for sound source localization (TDOA - Time Difference Of Arrival). The technical evaluation of the algorithm yielded an accuracy of about 5 cm, which seems entirely sufficient for our use case. A detailed description of the implementation and evaluation can be found in one of our previous publications [11]. The Leveler calculates the Euclidean distance between each microphone and the estimate source position and, based on these distances, determines gain values according to the 1/r law, which then are applied to the microphone signals. As a result, level changes caused by movements of the user are continuously compensated, and thus the reflections are presented at correct levels.

Regarding the diffuse part of the binaural room response, a specific reverberation source signal has to be acquired first. This processing step is handled by the ReSource filter module (ReSource - Reverberation Source), which is another part of the Processor. Ideally, the reverberation source signal can be obtained by superimposing all microphone signals. This sum signal, also called primal signal, contains all directional information of the source and therefore describes its full-spherical directivity. Summing all microphone signals in time domain, however, leads to interference artifacts caused by slight time differences between the signals, and thus is not feasible. Alternatively, the spectrum of the primal signal can be obtained by averaging the power spectra of all microphone signals. Indeed, this way all phase information gets lost, but as described in the following, only the spectrum of the primal signal is required in order to derive an appropriate excitation signal for the diffuse reverberation synthesis.

Briefly described, the ReSource algorithm continuously filters one single microphone signal so that it matches the spectrum of the primal signal. This filtered signal is then passed to the binaural renderer as the reverberation source signal. For this, the 32 microphone signals are permanently transformed to frequency domain with a short-time Fourier transform (STFT). For each frame of the STFT, the spectrum of the primal signal $S(n,k)$ is calculated by averaging the power spectra of all microphone signals, such as $S(n,k) = \sqrt{\sum_{i=1}^{N} w_i |(X_i(n,k)|^2}$,

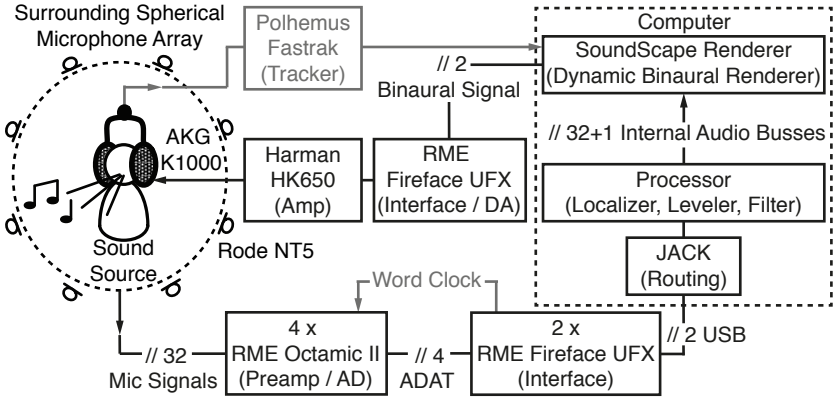


Fig. 5: Setup of the reactive VAE.

where n is the time index, k is the frequency bin, i is the microphone channel number, N is the number of microphones, $X_i(n,k)$ denotes the STFT of the ith microphone signal, and w_i indicates the weight of the ith microphone channel, which corresponds to the sphere surface area covered by the ith microphone (see segmentation per microphone in Sec. 3.2). Next, the spectra of the primal signal and of the microphone signal to be filtered are smoothed using logarithmic 1/3-octave spectrum smoothing. The power spectrum of the adaptive filter $F(n,k)$ is then calculated such as $F(n,k) = {S_{sm}(n,k)}/{|X_{i_{sm}}(n,k)|}$, where i is one specific microphone whose filtered signal will be used as the reverberation source signal. Here, one of the microphones located at the top of the array is used, mainly because level and spectrum changes caused by movements of the user are relatively small at the top microphones. After some regularization of $F(n,k)$, the minimum phase FIR filter $f(n)$ is derived from $F(n,k)$. Finally, the single-channel reverberation source signal is obtained by FFT-based uniform-partitioned convolution of $f(n)$ with the signal from microphone i.

The Processor has a separate process buffer independent of the system audio buffer. Generally, it works with any arbitrary process-buffer size, but we achieved the best results for the Localizer and the ReSource filter with a process-buffer size of 4096 samples. Thus, with a sampling rate of 48 kHz, the filter $f(n)$ and the levels are adapted approximately every 85 ms, which seems completely sufficient for our purpose. As informal tests showed, the filter resulting from a rotating sound source inside the array has a simple high-shelf character changing relatively slow over time, which indicates that the filter update rate can be quite moderate. Further tests with musicians revealed that the level update rate is also fast enough, as movements of the musicians are mostly quite smooth.

3.5 Setup

The complete setup of the reactive VAE is fairly straightforward (see Fig. 5). The 32 Rode NT5 array microphones are connected to four RME Octamic II preamps and AD converters, which again are connected to two RME Fireface UFX audio interfaces. Both interfaces work together as one aggregate device in the iMac computer. All further internal routing between the Processor, the SoundScape Renderer [12], and the aggregate device is realized with the JACK Audio Connection Kit. The signals are first routed through the Processor, which returns the level-adjusted microphone signals and the single-channel reverberation source signal. The 32 (leveled) microphone signals are then passed to the corresponding directional-BRIR sources in the renderer, and the reverberation source signal is sent to the diffuse-BRIR source. A Polhemus Fastrak provides the head-tracking data so that the renderer can generate the binaural signal according to the head orientation of the user. Finally, the binaural signal is DA converted with the main RME Fireface UFX interface, amplified with a Harman HK650 amplifier, and presented to the user over extra-aural AKG K1000 headphones.

To compensate for the magnitude response of the headphones and the microphones, specific compensation filters (minimum phase FIR filters) are applied to the BRIR datasets. Furthermore, to ensure correct level ratios between the user-generated sound and the synthesized room response, the system was level calibrated carefully (for more details, please refer to [7]).

4 Technical Evaluation

Similar to the implementation, the technical evaluation was an ongoing process. Besides checking every software and hardware component in detail, we successfully tested the basic functionality of the system in several ways and discussed systemic shortcomings [7]. Consequently, the technical evaluation in this paper mainly focuses on the Processor module, since it has only recently been implemented to finally complete the system.

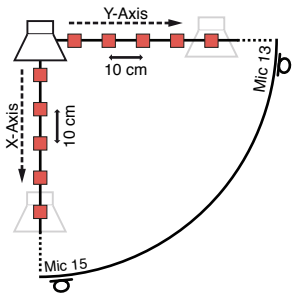


Fig. 6: Schematic top view of the array showing the measured source positions for the evaluation.

4.1 Leveler

The precision of the level adjustment largely depends on the accuracy of the source position estimation, which was determined as about 5 cm [11]. Furthermore, the sound source with its specific radiation pattern as well as the source signal have a distinct influence on the position estimation and on the level adjustment. To get an overview of how the Leveler performs in real time with various sound sources and source signals, we tested the algorithm for 11 different positions, 2 different sound sources, and 4 different source signals. As sound sources, we chose a Genelec 1029A loudspeaker and a HEAD acoustics HMSII.3 dummy head with the integrated mouth simulator. The latter was used to approximate a speaker directivity. The source signals were white noise, pink noise, speech, and guitar with a duration of 5 s each. As the schematic representation in Fig. 6 illustrates, we measured impulse responses of the array with the sources being placed along the x and y-axis at 6 positions each (shifts of 10 cm from the center of the array towards microphone no. 15 or no. 13 respectively). The sound sources were always oriented in direction of the x-axis. The captured impulse responses were then convolved with the source signals, resulting in a total of 96 multichannel audio streams. To assess the performance of the Leveler, we sent these streams through the Processor in real time, recorded the output of the module, and determined the RMS level of the recorded tracks.

Fig. 7 presents the results of the Leveler evaluation using the measurements with microphone no. 15 (a) and no. 13 (b) as examples. The plots show the differences between the RMS levels for the center position and the respective RMS levels for the shifted positions. Ideally, the levels should be identical, regardless of the source position, meaning that the level difference should always be around 0 dB. However, it seems that especially along the x-axis (see Fig. 7 (a)), the position estimation was quite unreliable, resulting in a relatively large variance of the level differences dependent on the source and the signal. In case of the Genelec loudspeaker, it is highly probable that its distinct source width as well as its orientation directly towards the microphone led to inaccurate position estimations and level adjustments. In comparison, the results are much better for the dummy head, quite likely because its source width is smaller. The maximum of the mean level differences (yellow line), however, is only about ± 2.5 dB, which seems to be in acceptable limits when compared to the mean level differences without the level adjustment (red dashed line), which are at about $+9$ dB for the x-axis shift of 50 cm.

As can be seen in Fig. 7 (b), the performance was much better for position shifts along the y-axis. Here, the variance dependent on sound source and source signal is significantly smaller, and thus the mean level differences provide a much more reliable measure. The mean level differences peak at about $+1.5$ dB for position shifts ≥ 30 cm, which seems quite good, especially when compared to the level differences which occur without level adjustments (about $+8$ dB for the y-axis shift of 50 cm).

Finally, using the results for the Genelec 1029A loudspeaker and the guitar source signal as an example, Fig. 8 illustrates the performance of the Leveler by means of level differences for all 32 microphones and the 6 positions along the y-axis. As can be seen, the variances around 0 dB are rather small if the levels are adjusted (blue triangles), resulting in a maximum standard deviation (yellow squares) of about 0.9 dB for the y-axis shift of 20 cm. Thus, depending on the estimated source position, all microphone signals are adequately increased or decreased in level in order to

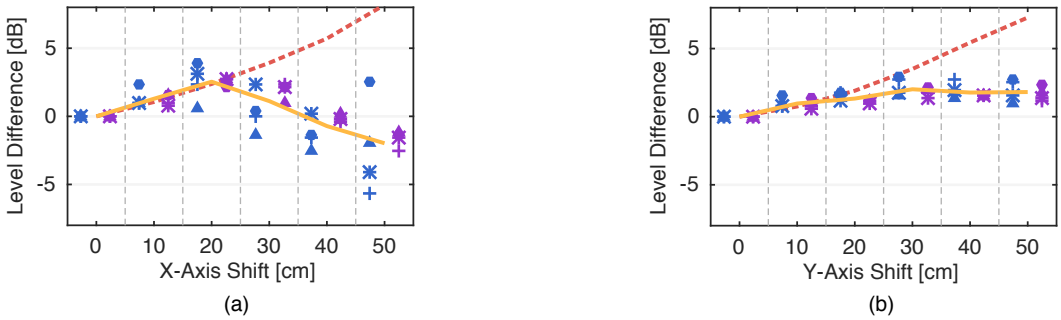


Fig. 7: Differences between the RMS levels for the center position and the respective RMS levels for the shifted position. Positions along the x (a) and y-axis (b), levels measured at microphone no. 13 (a) and no. 15 (b). Sound sources: Genelec 1029A (blue), HEAD acoustics HMSII.3 (purple). Source signals: white noise (asterisk), pink noise (plus sign), speech (circle), guitar (triangle). Yellow line: mean level difference per position. Red dashed line: mean level difference per position without level adjustments.

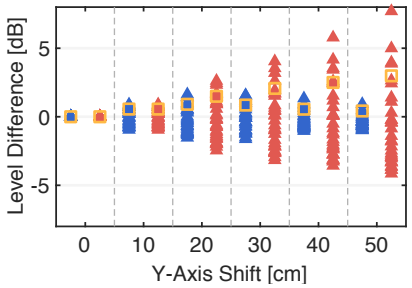


Fig. 8: Differences between the RMS levels for the center position and the respective RMS levels for the shifted position with (blue) and without (red) level adjustments for all 32 microphones and the positions along the y-axis. Sound source: Genelec 1029A. Source signal: guitar. Yellow square: standard deviation.

compensate for the off-center shift. Of course, without level adjustments, the variance of the level differences significantly increases as a function of the position shift (red triangles). In this example, the maximum standard deviation is about 4.3 dB for the y-axis shift of 50 cm. Overall, the Leveler performs quite well, even though there are some special cases where the position estimation (along the x-axis) provides imprecise results and the level adjustments are inaccurate. Nevertheless, the reactive VAE clearly benefits from the automatic level adjustment. Reflections are presented at adequate levels, no matter what the position of the sound source is, and the user can feel free to move while, for example, playing an instrument. Moreover, our informal tests showed that the user or the sound source usually stays quite centered most of the time. Thus, if the levels have to be adjusted at all, it is mostly only a slight adaption.

4.2 ReSource Filter

To evaluate the ReSource filter, we measured impulse responses of the array with a Genelec 1029A as the sound source. The loudspeaker was positioned in the center of the array and oriented in direction of the x-axis. Similar to the procedure described above, the captured impulse responses were convolved with a white noise signal (duration of 10 s) and sent through the Processor. The reverberation source signal, provided at the output of the Processor, was then recorded for further analysis. The corresponding primal signal, which is the reference in this case, was obtained by averaging the power spectra of the simulated microphone signals (the convolution results) as described in Sec. 3.4.
Fig. 9 shows the 1/3-octave smoothed magnitude spectrum of the reverberation source signal (red line), obtained by real-time processing with the ReSource filter, and of the actual primal signal (blue line). As indicated by the yellow dashed line, there are only negligible spectral differences, most likely induced by the spectral smoothing applied when designing the adaptive filter (see Sec. 3.4). Overall, the test demonstrates that the ReSource filter performs as expected and provides an appropriately filtered microphone signal at the output. Almost the same results were obtained for other loudspeaker orientations, which further confirms proper functioning of the implementation.

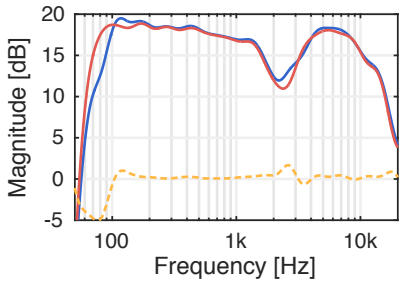


Fig. 9: Magnitude spectrum of the ReSource filter output (red line) and of the actual primal signal (blue line) as well as the difference between both spectra (yellow dashed line).

4.3 Latency

Depending on the BRIR length, the system runs smoothly with an audio buffer size of 128 or 256 samples ($f_s = 48$ kHz). With these settings, we measured a round-trip audio latency of about 14 or 22 ms respectively. The total audio latency is the result of many factors. For one thing, the sound propagation time from a source in the center of the array to the microphones results in about 3 ms of latency. The audio buffer (I/O) adds about 2 ms of latency for each 32 samples. The rest (about 3 ms) might be the sum of AD/DA converter latency and processing latency of the renderer. Yet, to a certain degree, the audio latency can be compensated, simply by shifting the BRIRs in time (see Sec. 3.2 and 3.3). As an example, if the system runs with a buffer size of 128 samples and the floor reflection is omitted (as outlined in Sec. 3.2), the audio latency can be fully compensated for rooms with a minimum distance of about 2.50 m between the source and the walls.
Besides the round-trip audio latency, there is also a system latency in dynamic binaural synthesis that is defined as the delay between a head movement and the corresponding reaction of the VAE, like rendering the binaural signal with updated BRIRs. Similar as for the audio latency, different elements contribute to the total system latency, such as the update rate of the head tracker, the audio buffer size, and the processing time of the renderer. In our case, we measured a system latency of about 32 and 40 ms with a head tracker update rate of 120 Hz and an audio buffer size of 128 or 256 samples respectively ($f_s = 48$ kHz). This is sufficiently below assessed thresholds of just detectable system latency of about 60 - 70 ms [13].

5 Conclusion

In this paper, we presented a headphone-based reactive VAE that allows sonic interaction of the user with a virtual environment through self-generated sound. As the system provides new opportunities for natural interaction with the virtual room, potential applications in the field of immersive audio are diverse. Examples are virtual performance spaces as well as any interactive virtual-reality experience. In our case, the reactive VAE mostly serves as a research tool to conduct experiments on the perception of self-generated sound in rooms.
The system is based on a 32-channel surrounding spherical microphone array to capture the direction-dependent sound of the user. In comparison to other reactive VAEs introduced so far, this setup has two clear advantages: it generally works with any type of sound source, and the dynamic directivity of the source is adequately considered in the reproduction. For this, the synthesized BRIRs, describing the direction-dependent reflections (directional BRIRs) and the binaural reverberation (diffuse BRIR) of the room are important. The parameter-based synthesis of the BRIRs allows varying their properties systematically, which is also beneficial for perceptual studies.
To complete the technical evaluation of the overall system, we presented several tests of the Processor, a real-time software module that automatically levels the microphone signals dependent on the source position (Localizer and Leveler) and generates the reverberation source signal (ReSource filter). The tests revealed some special cases where the Localizer and the Leveler provide inaccurate results, but overall the system works correctly and greatly benefits from the automatic level adjustment. The ReSource filter, however, proved to perform very well.
In future work, we plan to use the reactive VAE as a tool for psychoacoustic experiments. For example, we want to investigate to what extent users are able to perceive a change in spatial resolution of the capturing system. Thus, it may be possible that the number of microphones can be significantly decreased without notable perceptual degradation. If true, a greatly simplified version of the system could be implemented, for example based on a small number of clip-on microphones. In combination with simulation or parametric synthesis of the directional BRIRs in real time, a system could be realized that allows the user to move freely in the virtual room. Furthermore, we want to examine echo thresholds for self-generated sound in compar-

ison to external sound sources. The masking effect of self-generated sound will most probably increase the thresholds. Finally, we plan to further investigate the influence of self-generated sound on immersion and presence. Previous studies already revealed significantly increased presence in reactive VAEs compared to usual reproduction systems.
However, established psychoacoustic test procedures are mostly unsuitable for experiments with self-generated sound. For example, there is simply no pre-recorded test signal that can be played back. Thus, the question arises how to obtain a reproducible test signal, meaning what kind of sound the subject has to generate, and in which way the sound has to be generated. A reasonable solution seems to be that participants have to read a prepared text, or that they have to play an easy instrument like a snare drum to excite the virtual room. Yet, it might be that participants have to train the sound generation if repeatability and reproducibility is required. Another problem is that in a properly calibrated system, the room response is rather quiet compared to the direct sound. Hence, if the natural level ratio between direct and reflected sound is maintained, it is extremely difficult for participants to hear small variances in the synthesized room response. Furthermore, paying attention to self-generated sound might be an unusual task, which is why specific training sessions might be required. Even more general, the question arises how participants control the experiment and submit their ratings if they have to play an instrument, simply because they have no free hands to hold a controlling device, and because they cannot wear a head-mounted display as they have to see the instrument while playing. Thus, it becomes apparent that new concepts have to be developed first in order to conduct reproducible studies on the perception of self-generated sound. This in turn opens up new research possibilities in the field of sonic interaction.

References

[1] Laird, I., Murphy, D. T., Chapman, P., and Jouan, S., "Development of a Virtual Performance Studio with application of Virtual Acoustic Recording Methods," in *Proc. 130th AES Conv., London, UK*, pp. 1–12, 2011.

[2] Schärer Kalkandjiev, Z., *The Influence of Room Acoustics on Solo Music Performances. An Empirical Investigation*, Doctoral dissertation, TU Berlin, 2015.

[3] Pörschmann, C., "One's Own Voice in Auditory Virtual Environments," *Acta Acust. united Ac.*, 87(3), pp. 378–388, 2001.

[4] Flanagin, V. L., Schörnich, S., Schranner, M., Hummel, N., Wallmeier, L., Wahlberg, M., Stephan, T., and Wiegrebe, L., "Human Exploration of Enclosed Spaces through Echolocation," *J. Neurosci.*, 37(6), pp. 1614–1627, 2017.

[5] Postma, B. N., Demontis, H., and Katz, B. F. G., "Subjective Evaluation of Dynamic Voice Directivity for Auralizations," *Acta Acust. united Ac.*, 103(2), pp. 181–184, 2017.

[6] Vigeant, M. C. and Wang, L. M., "Investigations of Orchestra Auralizations Using the Multi-Channel Multi-Source Auralization Technique," *Acta Acust. united Ac.*, 94(6), pp. 866–882, 2008.

[7] Arend, J. M., Stade, P., and Pörschmann, C., "Binaural reproduction of self-generated sound in virtual acoustic environments," *Proc. Meetings on Acoustics*, 30(1), pp. 1–14, 2017.

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

[9] Bernschütz, B., "A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100," in *Proc. 39th DAGA*, pp. 592–595, 2013.

[10] Stade, P., Arend, J. M., and Pörschmann, C., "Perceptual Evaluation of Synthetic Early Binaural Room Impulse Responses Based on a Parametric Model," in *Proc. 142nd AES Conv., Berlin, Germany*, pp. 1–10, 2017.

[11] Lübeck, T., Arend, J. M., and Pörschmann, C., "A Real-Time Application for Sound Source Localization Inside a Spherical Microphone Array," in *Proc. 44th DAGA*, pp. 319–322, 2018.

[12] Geier, M., Ahrens, J., and Spors, S., "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods," in *Proc. 124th AES Conv., Amsterdam, The Netherlands*, pp. 1–6, 2008.

[13] Lindau, A., "The Perception of System Latency in Dynamic Binaural Synthesis," in *Proc. 35th DAGA*, pp. 1063–1066, 2009.

3.4 Binauralization of Omnidirectional Room Impulse Responses – Algorithm and Technical Evaluation

Pörschmann, C., Stade, P., & Arend, J. M. (2017). In *Proc. of the 20th International Conference on Digital Audio Effects (DAFx17), Edinburgh, UK* (pp. 345–352).

BINAURALIZATION OF OMNIDIRECTIONAL ROOM IMPULSE RESPONSES - ALGORITHM AND TECHNICAL EVALUATION

Christoph Pörschmann [*], *Philipp Stade* [*] [†], *Johannes M. Arend* [*] [†]

[*]TH Köln, Institute of Communication Engineering, Cologne, Germany
[†]TU Berlin, Audio Communication Group, Berlin, Germany
christoph.poerschmann@th-koeln.de

ABSTRACT

The auralization of acoustic environments over headphones is often realized with data-based dynamic binaural synthesis. The required binaural room impulse responses (BRIRs) for the convolution process can be acquired by performing measurements with an artificial head for different head orientations and positions. This procedure is rather costly and therefore not always feasible in practice. Because a plausible representation is sufficient for many practical applications, a simpler approach is of interest.
In this paper we present the *BinRIR* (Binauralization of omnidirectional room impulse responses) algorithm, which synthesizes BRIR datasets for dynamic auralization based on a single measured omnidirectional room impulse response (RIR). Direct sound, early reflections, and diffuse reverberation are extracted from the omnidirectional RIR and are separately spatialized. Spatial information is added according to assumptions about the room geometry and on typical properties of diffuse reverberation. The early part of the RIR is described by a parametric model and can easily be modified and adapted. Thus the approach can even be enhanced by considering modifications of the listener position. The late reverberation part is synthesized using binaural noise, which is adapted to the energy decay curve of the measured RIR.
In order to examine differences between measured and synthesized BRIRs, we performed a technical evaluation for two rooms. Measured BRIRs are compared to synthesized BRIRs and thus we analyzed the inaccuracies of the proposed algorithm.

1. INTRODUCTION

Binaural synthesis is a powerful tool for headphone-based presentation of virtual acoustic environments (VAEs). It can be applied for auralization purposes in various areas like audio engineering, telecommunication, or architectural acoustics. For many of these applications, a plausible presentation is sufficient; an authenticity reproduction of the sound field is often not pursued. In this context plausibility refers to the illusion that the scenario being depicted is actually occurring [1] while authentic refers to a perception that the scenario cannot be distinguished from a real reference.
Binaural room impulse responses (BRIRs) can be applied, which are either simulated or measured with an artificial head for different orientations (and positions). Finally the BRIRs are convolved with anechoic signals in a binaural renderer. By considering the listener's head movements in the auralization, localization accuracy increases [2], front-back confusion can be decreased [2] and externalization of virtual sound sources improves [3][4]. Several commercial or scientific rendering engines are available, which adapt the sound field presented through headphones according to the orientation of the listener in real time (e.g. [5][6][7][8]). Depending on the head movements, which shall be considered in the auralization, these measurements need to be done for circular orientations in the horizontal plane or even for spherical head orientations considering horizontal and vertical rotations. However, measuring such BRIR datasets requires a large amount of time and the use of complex devices (e.g. a rotatable artificial head). Furthermore, for each listening position, another set of BRIRs needs to be captured. Thus, for many applications in the field of spatial audio and virtual environments, the effort is so high that circular sets of BRIRs are not used. To approach this issue, we developed the *BinRIR* (Binauralization of omnidirectional room impulse responses) algorithm, which aims for an auralization based on a simple measurement procedure. Only one single measured omnidirectional room impulse response (RIR) is required to obtain a plausible auralization when using dynamic binaural synthesis. The algorithm even allows to shift the listener position. Thus, one single measured RIR is sufficient to synthesize a BRIR dataset for a freely chosen head orientation and position in the room.
In literature, several approaches to obtain BRIRs from measured RIRs have been described. In [9][10] a synthesis of BRIRs from B-format measurements has been proposed. The spatial impulse response rendering (SIRR) method applies a decomposition of the sound field into direct and diffuse parts. While the diffuse part is decorrelated, vector-based amplitude panning is used to distribute the direct sound on different loudspeakers. In [11] a directional audio coding (DirAC) method is proposed which can capture, code, and resynthesize spatial sound fields. DirAC analyzes the audio signal in short time frames and determines the spectrum together with direction and diffuseness in the frequency bands of human hearing. As this method does not work impulse response-based it is quite different to the one presented in this paper. Another simple approach to synthesize BRIRs has been presented by Menzer. In [12][13] RIRs measured in the B-format are used to synthesize BRIRs. Direct sound, spectral shape and temporal structure are extracted from the RIR. Additionally, the incidence direction of the direct sound is estimated from the measured data. No specific treatment of the early reflections is proposed. All reflections and the diffuse reverberation are synthesized by performing an adequate reconstruction of the interaural coherence.
In this paper, we present research results on the binauralization of omnidirectionally measured RIRs. Parts of the studies including the basic idea and a basic description of the approach as well as results of a perceptual evaluation have already been published [14][15][16]. This paper is organized as follows: In section 2 we introduce and explain the *BinRIR* algorithm performing the spatialization of omnidirectional RIRs in detail. In section 3 we describe the results of a technical evaluation. We compare measured BRIRs of two different rooms to the synthesized counterparts and elaborate differences caused by the simplifications of the algorithm. Finally, section 4 concludes the paper and provides an outlook.

2. ALGORITHM DESIGN

2.1. General Structure

The basic idea of the *BinRIR* algorithm is to use only one measured omnidirectional RIR for the synthesis of BRIR datasets which can be used for dynamic auralization. The algorithm was implemented in Matlab and applies predictable information from sound propagation in enclosed spaces as well as knowledge regarding the human perception of diffuse sound fields. For processing, the RIR is split into different parts. The early part contains the direct sound and strong early reflections. For this part, the directions of incidence are modeled reaching the listener from arbitrarily chosen directions. The late part of the RIR is considered being diffuse and is synthesized by convolving binaural noise with small sections of the omnidirectional RIR. By this, the properties of diffuse reverberation are approximated. The algorithm includes an additional enhancement: The synthesized BRIRs can be adapted to shifts of the listener and thus freely chosen positions in the virtual room can be auralized.

The *BinRIR* algorithm incorporates several inaccuracies and deviates significantly from a measured BRIR. The directions of incidence of the synthesized early reflections are not in line with the real ones. Hence, differences in the perception of spatial properties (e.g. envelopment) between the original room and the synthesized room may occur. Furthermore, a point source is assumed for all synthetic BRIR datasets. Thus, it is not possible to rebuild source width and other properties of the source correctly. Finally, the diffusely reflected part of the early reflections cannot be precisely reconstructed.

The basic structure of the *BinRIR* algorithm is shown in Figure 1. As input data the algorithm only requires the omnidirectional RIR and the position of the sound source. Furthermore, the algorithm accesses an appropriate set of HRIRs and a preprocessed sequence of binaural noise. Both were obtained from measurements with a Neumann KU100 artificial head [17].

The algorithm is only applied to frequencies above 200 Hz. For lower frequencies the interaural coherence of a typical BRIR is nearly one and the omnidirectional RIR can be maintained. 7th order Chebyshev Type II filters are used to separate the low frequency part from the rest of the signal.

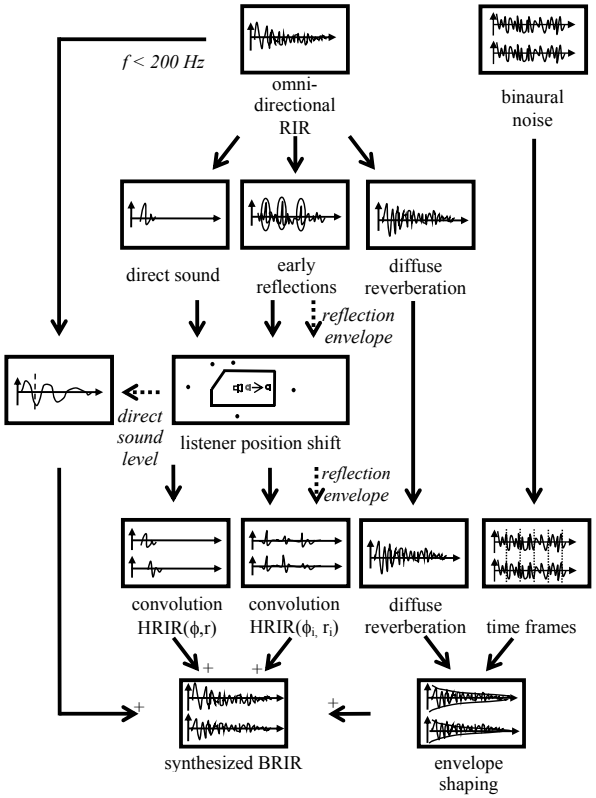


Figure 1: Block diagram of the *BinRIR* algorithm for synthesizing a BRIR based on one single omnidirectional RIR.

2.2. Direct sound and early reflections

Onset detection is used to identify the direct sound in the omnidirectional RIR. The direct sound frame starts with the onset and ends after 10 ms (5 ms followed by 5 ms raised cosine offset). The following time section is assigned to the early reflections and the transition towards the diffuse reverberation. For small and non-reverberant rooms ($V < 200\,m^3$ and $RT_{60} < 0.5$ s) a section length of 50 ms is chosen, otherwise the section is extended to 150 ms. In order to determine sections with strong early reflections in the omnidirectional RIR, the energy is calculated in a sliding window of 8 ms length and time sections which contain high energy are marked. Peaks which are 6 dB above the RMS level of the sliding window are determined and assigned to geometric reflections. A windowed section (raised cosine, 5 ms ramp) around each of the peaks is considered as one reflection. If several very dense reflections occur in adjacent sections, these sections are merged. Following this procedure, small windowed sections of the omnidirectional RIR are extracted describing the early reflections. The incidence directions of the synthesized reflections base on a spatial reflection pattern adapted from a shoebox room with non-symmetric positioned source and receiver. Thus a fixed lookup-table containing the incidence directions is used. By this a simple parametric model of the direct sound and the early reflections is created. Amplitude, incidence direction, delay and the envelope of each of the reflections are stored. The design of the algorithm identifying the reflections and its parameterization (e.g. window length, peak detection) was done based on empiric tests. Informal listening experiments during the development have shown that the exact way in which the reflections are determined is not substantial.

By convolving each windowed section of the RIR with the HRIR(φ) of each of the directions, a binaural representation of the early geometric reflective part is obtained. To synthesize interim directions between the given HRIRs, interpolation in the spherical domain is performed [18].

2.3. Diffuse Reverberation

The diffuse reverberation is considered reaching the listener temporarily and spatially equally distributed. Several studies have shown that after a so-called perceptual mixing time, no specific differences to completely diffuse reverberation can be perceived and thus, an exact reconstruction of the temporal and spatial structure is not required (e.g. [19]). It was found that the perceptual mixing time is room dependent and can be chosen according to predictors which are calculated based on geometric room prop-

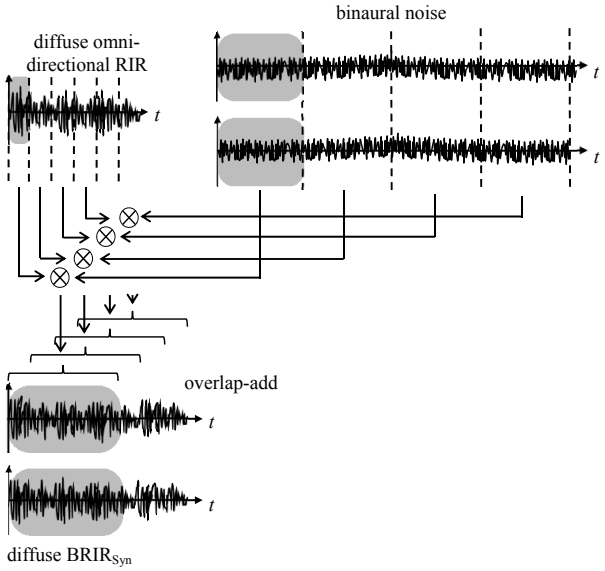


Figure 2: Synthesis of the binaural diffuse reverberation: Sections of the diffuse omnidirectional RIR (0.67 ms; 32 taps at 48 kHz sampling rate) and the binaural noise (2.67 ms; 128 taps at 48 kHz sampling rate) are convolved. Both sections are raised-cosine windowed. The diffuse BRIR is synthesized by summing up the results of the convolutions applying overlap-add.

erties. However, in [20] it has been shown that small perceptual differences still remain. Thus recent studies (e.g. [21][22]) proposed models applying an overlap between early reflections and the diffuse part instead of using a fixed mixing time. By this some diffuse energy is embedded in the early part of the BRIR. A similar approach is used in the *BinRIR* algorithm: All parts of the RIR excluding the sections of the direct sound and the detected early reflections are assigned to the diffuse part.

To synthesize the binaural diffuse reverberation we developed two different methods. In [14][15] the RIR was split up into 1/6 octave bands by applying a near perfect reconstruction filter bank [23] and the binaural diffuse part was synthesized for each frequency band. In [16] we proposed another method for the synthesis of the diffuse part which is used in this publication. The diffuse reverberation is synthesized by convolving small time sections (0.67 ms) of the omnidirectional RIR with sections of 2.67 ms binaural noise (both sections raised-cosine windowed). The results of the convolutions of all time sections are summed up with overlap-and-add. Figure 2 explains the synthesis of the diffuse reverberation in greater detail. By this, both the binaural features (e.g. interaural coherence) of the binaural noise and the frequency-dependent envelope of the omnidirectional RIR are maintained. The lengths of the time sections were determined by informal listening tests during the development of the algorithm. This method requires less computational power than the one proposed in [14][15]. Informal listening tests showed that both methods are perceptually comparable.

2.4. Listener position shifts

The algorithm includes a further enhancement: The synthesized BRIR can be adapted to listener position shifts (LPS) and thus freely chosen positions of the listener in the virtual room can be

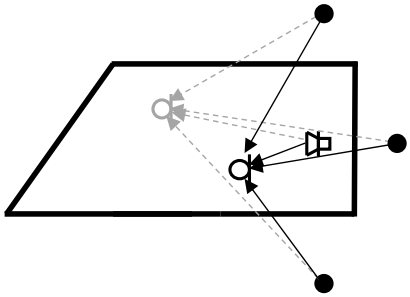


Figure 3: Basic principle of the listener position shifts (LPS): A mirror-image model is applied to modify the amplitude and the temporal structure of the direct sound and the early reflections. The receiver is moved from an initial position (grey) to a modified position (black). By this the paths of the direct sound and the reflections are changed.

auralized. For this, a simple geometric model based on mirror-image sound sources is used. The distance between the listener and each of the mirror-image sound sources is determined from the delay of the corresponding reflection peak to the direct sound peak. In a next step, a shifted position of the listener is considered and amplitudes (based on the 1/r law), distances, and directions of incidence are recalculated for each reflection (Fig. 3). Optimizing an earlier version of the *BinRIR* algorithm (e.g [16]) we modified the low-frequency component below 200 Hz when applying LPS. If, for example, the listener approaches the sound source, the amplitude of the direct sound increases and the low-frequency energy for the direct sound needs to be adapted accordingly. For this the low-frequency part of the direct sound (first 10 ms followed by 10 ms raised cosine set) is adjusted according to the 1/r law.

2.5. Synthesis of Circular BRIR sets

The synthesis of the BRIRs is repeated for constant shifts in the azimuthal angle (e.g. 1°) for the direct and the reflected sound. Thus, a circular set of BRIRs is obtained, which can be applied by different rendering engines for dynamic binaural synthesis. The synthesized sets of BRIRs can be stored in various formats, e.g the miro-Format [24], a multi-channel-wave format to be used by the SoundScape Renderer [7] and can be converted to the SOFA-format [25].

3. TECHNICAL EVALUATION

To analyze the performance of the algorithm, a series of listening experiments has already been conducted [14][16]. These experiments mainly aimed at a quantification of the perceptual differences between measured and synthesized BRIRs. In this paper we focus on a technical evaluation of the algorithm and compare different properties of the synthesized BRIRs to the properties of measured BRIRs. Therefore we analyzed the detected reflections regarding their direction, their time of incidence, and their amplitude to measured data. Furthermore, we compared the reverberation tails and the reverberation times of the synthesized BRIRs to the ones of measured BRIRs. We investigated to what extent the clarity (C_{50}) of the synthesized room matches the measured room's

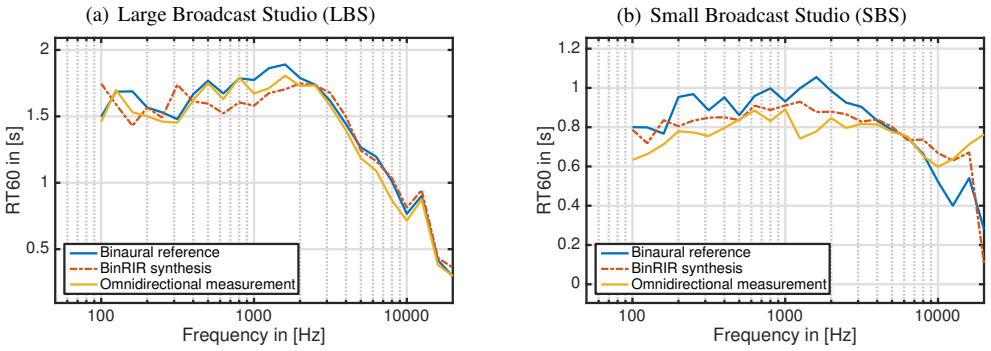


Figure 4: Reverberation Time (RT_{60}) of the Large Broadcast Studio (a) and the the Small Broadcast Studio (b). In each plot the RT_{60} for the binaurally measured reference, for the synthesis with the *BinRIR* algorithm and for the omnidirectional measurement are shown. The RT_{60} was calculated in 1/3 octave bands in the time domain

clarity. Finally we looked briefly in which way the use of the LPS influences the early part of the BRIR.

3.1. Measured rooms

The performance of the algorithm was analyzed for two different rooms. Both rooms are located at the WDR radiobroadcast studio in Cologne and are used for various recordings of concerts and performances. The "*KVB-Saal*" (Large Broadcast Studio - LBS) has a volume of 6100 m^3, a base area of 579 m^2 and can seat up to 637 persons. We measured the impulse responses in the 6th row ($Distance_{SrcRec} = 13.0$ m). The "*kleiner Sendesaal*" (Small Broadcast Studio - SBS) has a volume of 1247 m^3, a base area of 220 m^2 and 160 seats. The $Distance_{SrcRec}$ in this room was 7.0 m. In order to evaluate the algorithm, measured impulse responses from these rooms were used. In addition to the omnidirectional RIRs, which are required to feed the *BinRIR* algorithm, we measured circular BRIR datasets at the same position as a reference. This dataset was measured in steps of 1° on the horizontal plane with a Neumann KU100 artificial head. Finally we used data from spherical microphone array measurements, conducted with the VariSphear measurement system [26]. For this, we applied a rigid sphere array configuration with 1202 sample points on a Lebedev grid at a diameter of 17.5 cm. The omnidirectional RIRs and the array data were measured with an Earthworks M30 microphone. As sound source, a PA stack involving an AD Systems Stium Mid/High unit combined with 3 AD Systems Flex 15 subwoofers was used. The complete series of measurements is described in detail in [27].
Based on the microphone array measurements, we identified reflections in the room using sound field analysis techniques [28]. For this the array impulse responses were temporally segmented (time resolution = 0.5 ms) and transformed into the frequency domain. Applying a spatial Fourier transformation, the impulse responses were transformed into the spherical wave spectrum domain [29]. Then the sound field was decomposed into multiple plane waves using the respective spatial Fourier coefficients. Data was extracted for a decomposition order of N = 5, a spherical composite grid with 3074 Lebedev points at a frequency f = 4500 Hz. For this frequency quite robust results for the detection of the reflections were found. By this, a spatio-temporal intensity matrix of the sound field at the listener position was calculated. Each time slice of the matrix was analyzed with a specific algorithm in order to classify reflections which are represented as local maxima in the matrix [30]. The detected reflections were stored with their attributes "time", "direction" and "level" in a reflection list and can be used for further comparison with the *BinRIR* algorithm.

3.2. Direct Sound and early reflections

In a first step we looked at the early part of the synthetic BRIRs and compared the direct sound and the early reflections determined by the *BinRIR* algorithm to the reflections which are identified using sound field analysis techniques based on the microphone array measurements. To describe the directions of the reflections we used a spherical coordinate system. The azimuthal angle denotes the orientation on the horizontal plane with $\varphi = 0°$ corresponding to the front direction. The elevation angle is $\delta = 0°$ for frontal orientation and $\delta = 90°$ for sound incidence from above.
Table 1 and 2 show the reflections detected by *BinRIR* as well as the 14 strongest reflections which were identified based on the array data. For each reflection, the time of arrival relative to the direct sound, the incidence direction, and the level are shown. The temporal structure and the energy of many of the reflections show similarities. For example in the LBS for reflection # 7, # 10 and # 13 and in the SBS for reflection # 13 the level and the time of arrival match quite well. Furthermore, some of the reflections were detected in the array measurements as impinging from different directions nearly simultaneously. These reflections are typically merged to one reflection with an increased level by the *BinRIR* algorithm (e.g. LBS: # 4 and # 5; SBS: # 6 and # 7; # 9 - # 11).
As the incidence directions of the synthetic reflections are chosen by *BinRIR* based on a lookup-table, it is not surprising that there are significant differences compared to the directions of the measured reflections. However, if the proportions as well as source and listener position of the synthesized room to some extent match the modelled shoebox room some of the incidence directions are appropriate. Thus, at least for the first reflection and as well for some other reflections detected by *BinRIR*, an acceptable congruence exists both for the LBS and the SBS. The azimuthal deviation of the first reflection is less than 40° and in elevation no relevant deviation exists.
As already explained in 2.2 the *BinRIR* algorithm starts detecting reflections 10 ms after the direct sound. Thus no reflections reaching the listener within the first 10 ms can be found. The time frame

	Array measurement				BinRIR algorithm			
#	Delay [ms]	Azimuth [°]	Elevation [°]	Level [dB]	Delay [ms]	Azimuth [°]	Elevation [°]	Level [dB]
0	0.0	0	3	0.0	0.0	0	0	0.0
1	5.0	8	-58	-27.6				
2	7.5	142	-18	-27.5				
3	17.5	0	-8	-27.3				
4	24.0	309	0	-25.6	24.9	319	-2	-21.0
5	26.5	58	5	-27.9				
6	28.0	278	0	-28.1				
7	31.0	0	58	-20.3	31.1	10	-2	-19.4
8	38.0	253	-2	-28.0	38.4	85	-3	-22.9
9	50.5	180	20	-27.2				
10	79.5	178	17	-12.7	79.6	343	4	-14.9
11	92.5	353	73	-26.6	92.5	319	-54	-40.5
12					108.2	0	67	-34.4
13	117.5	183	35	-26.3	117.6	85	-50	-31.4
14	174.0	356	8	-25.0				
15	202.0	3	3	-24.3				

Table 1: Properties of the direct sound and the early reflections for the Large Broadcast Studio (LBS). Left side: Reflections determined from the analysis of the array data. Right side: Reflections detected by the *BinRIR* algorithm

	Array measurement				BinRIR algorithm			
#	Delay [ms]	Azimuth [°]	Elevation [°]	Level [dB]	Delay [ms]	Azimuth [°]	Elevation [°]	Level [dB]
0	0.0	5	5	0.0	0.0	0	0	0.0
1	1.5	7	-22	-6.9				
2	3.5	16	-76	-21.9				
3	7.5	210	-22	-27.5				
4	12.0	204	-17	-27.7				
5	15.5	27	66	-23.9				
6	20.0	284	0	-21.1	22.2	319	-2	-19.2
7	23.0	58	75	-24.5				
8	35.0	203	2	-27.7	35.4	10	-2	-21.6
9	50.0	152	1	-21.5				
10	50.5	151	-2	-21.5	51.3	85	-3	-18.5
11	51.5	147	29	-25.5				
12	57.0	129	0	-21.4				
13	59.0	173	7	-19.7	59.6	343	4	-15.9
14	99.5	331	-2	-19.5	100.3	319	-54	-29.3

Table 2: Properties of the direct sound and the early reflections for the Small Broadcast Studio (SBS). Left side: Reflections determined from the analysis of the array data. Right side: Reflections detected by the *BinRIR* algorithm

determining the geometric reflections ends after 150 ms and no reflections are detected by *BinRIR* after this period.
Furthermore, several reflections which can be extracted from the array data measurements are not detected by *BinRIR* (e.g. # 3 and # 9 in the LBS and # 5 and # 12 in the SBS). However, in total more than 2/3 of the reflections in the section from 10 ms to 150 ms determined from the array data correspond to a reflection determined by *BinRIR*.

3.3. Energy Decay and Reverberation Time

In a next step we compared the energy decay curves of the synthesis and of the binaurally measured reference BRIRs. As already explained, the *BinRIR* algorithm synthesizes diffuse reverberation by applying a frame-based convolution of the omnidirectional RIR with binaural noise. Thus in addition to the synthesized and the measured BRIR (reference) we analyzed the energy decay of the measured omnidirectional RIR as well. The following analysis is based on the impulse responses for the frontal viewing direction ($\varphi = 0°$, $\delta = 0°$). Analyzing the reverberation time RT_{60} (Figure 4) we observed that the general structure of the curves is similar, but variations between the three curves exist. The average unsigned deviation between the synthesis and the reference is 0.10 s for the LBS and 0.09 s for the SBS, the maxima are 0.26 s (LBS) and 0.23 s (SBS).

3.4. Interaural Coherence

Next, we compared the interaural coherence (IC) of the synthesized BRIRs and of the reference BRIRs (Figure 5). We calculated the IC according to [31] applying hamming-windowed blocks with a length of 256 taps (5.33 ms) and an overlap of 128 taps (2.67 ms). In each plot the IC calculated with three different starting points is shown. For reference and synthesis in both rooms the IC is significantly different when direct sound is included in the calculation (t > 0 ms). For the medium condition (LBS: t > 150 ms; SBS: t >50 ms) significant differences between synthesis and reference can be observed. This is not surprising as no shaping of the IC is performed in the *BinRIR* algorithm. However, this difference is smaller for the SBS, because the impulse response is probably nearly diffuse at 50 ms. For the condition with the maximal starting point (LBS: t > 300 ms; SBS: t > 150 ms) which mainly comprises the diffuse reverberation the IC of the synthesized BRIR matches the reference quite well.

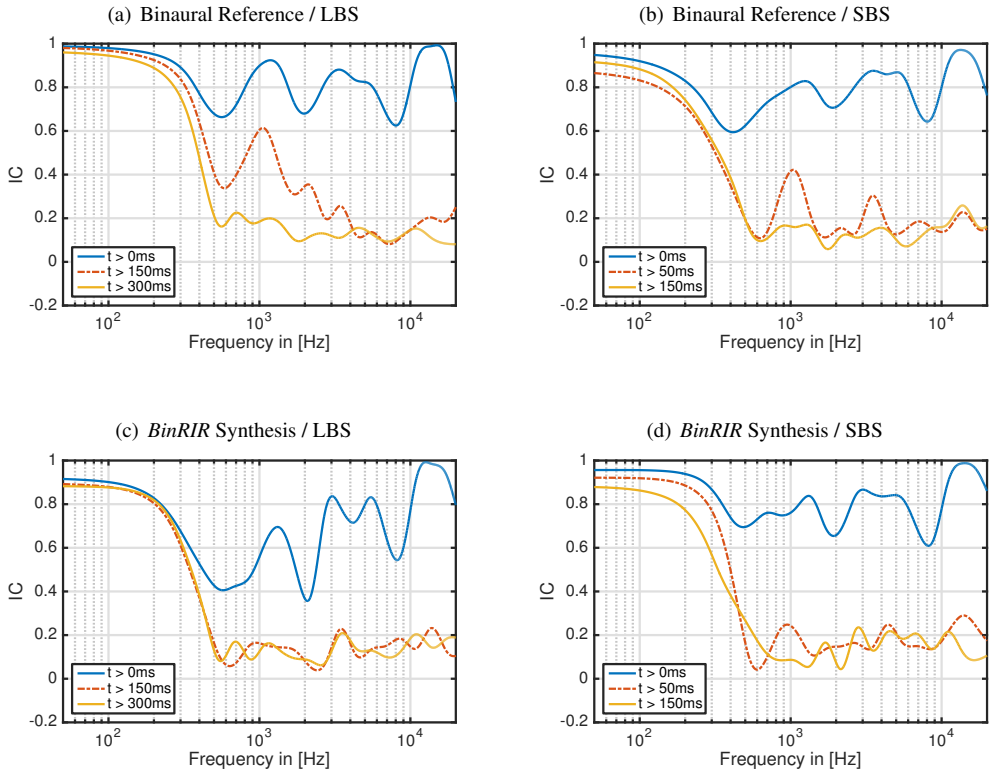


Figure 5: Interaural coherence (IC) of the Large Broadcast Studio (LBS) and the the Small Broadcast Studio (SBS). In plot (a) and (b) the data for the binaural reference is shown, in (c) and (d) the data for the *BinRIR* synthesis. For the LBS the IC is plotted for t > 0 ms, t > 150 ms and t >300 ms, for the SBS for t > 0 ms, t > 50 ms and t >150 ms.

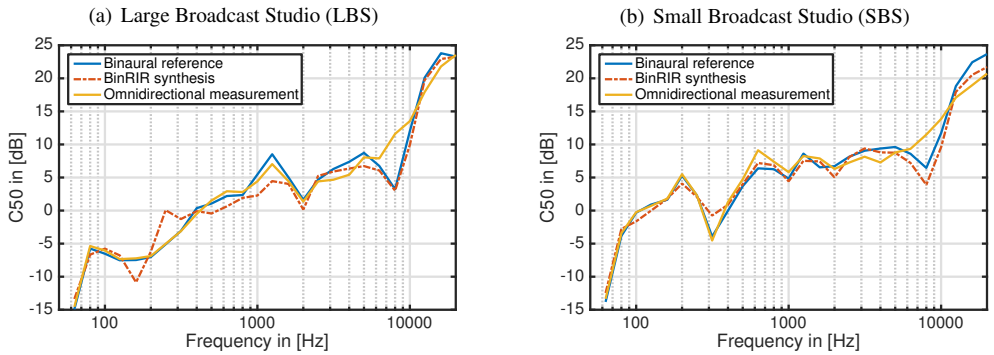


Figure 6: Clarity (C_{50}) of the Large Broadcast Studio (LBS) and the the Small Broadcast Studio (SBS). In each plot the C_{50} for the binaurally measured reference, for the synthesis with the *BinRIR* algorithm and for the omnidirectional measurement are shown.

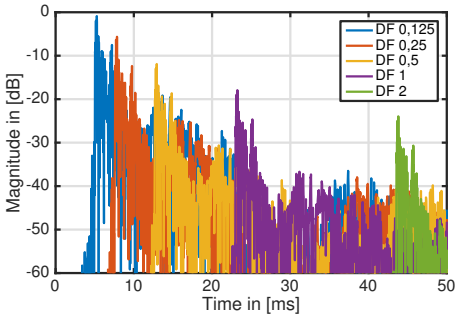


Figure 7: Influence of the Listener Position Shift (LPS) on the early part of the time response for the Small Broadcast Studio (SBS). The time responses for Distance Factors (DFs) from 0.125 - 2 are shown in different colors

3.5. Clarity

Next, we examined the clarity C_{50} over frequency for each of the conditions (Figure 6). The differences between the omnidirectional RIR, the synthesis and the reference are minor. The average unsigned deviation between the synthesis and the reference is 1.4 dB for the LBS and 1.1 dB for the SBS. The maxima are 5.2 dB (LBS) and 3.2 dB (SBS). Thus the ratio of the energy of the early part of the BRIR and the late diffuse part of the BRIR can be regarded as appropriate.

3.6. Listener position shifts

Finally we analyzed for the SBS in which way the early part of the synthesized BRIR is changed when listener position shifts (LPS) are performed. The results are shown in Figure 7 for synthesized $\text{Distance}_{\text{SrcRec}}$ between 0.875 m and 14 m (distance factor 1/8 - 2 of the original distance). It can be observed that the level and the time of arrival of the direct sound are changed significantly (according to the 1/r distance law) when performing LPS. The influence of the LPS on reflections is hard to observe from the plot, but changes in amplitude and time of arrival according to the geometric room model can be found here as well. The diffuse part and thus the complete late reverberation remain unchanged when performing LPS (not shown in Figure 7).

4. CONCLUSION

In this paper, the *BinRIR* algorithm was presented, which aims for a plausible dynamic binaural synthesis based on one measured omnidirectional RIR. In two different rooms, RIRs were measured and binauralized applying the presented *BinRIR* algorithm, so that synthetic BRIR datasets were generated. The presented method separately treats direct sound, reflections and diffuse reverberation. The early parts of the impulse responses are convolved with HRIRs of arbitrary chosen directions while the reverberation tail is rebuilt from an appropriately shaped binaural noise sequence. In an extension, the algorithm allows to modify the sound source distance of a measured RIR by changing several parameters of an underlying simple room acoustic model.
The synthetic BRIRs were compared to reference BRIRs measured with an artificial head. Due to missing information on spatial aspects, a perfect reconstruction of the sound field is generally not possible. An analysis of the early reflections showed that neither all reflections are detected by the *BinRIR* algorithm nor their directions match to the physical ones of the room. However, the reflections which were identified by the *BinRIR* algorithm correlate with the times of incidence and partly with the direction of incidence of the physical reflections in the room quite well. For the diffuse part, small differences in the reverberation time and the interaural coherence were observed. However, in general, the synthesis can be regarded as appropriate. An evaluation of the reverberation time RT_{60} and of the clarity C_{50} only showed minor differences between reference and synthesis. Analyzing the perceptual influences of the determined differences is not covered in the study presented here. Please refer to [14][16] for an analysis of these topics.
The approach presented in this paper can be combined with other modifications of measured RIRs. In [32][33], we discussed a predictive auralization of room modifications by an appropriate adaptation of the BRIRs. Thus the measurement of one single RIR is sufficient to obtain a plausible representation of the modified room. Furthermore, the opportunity to shift the listener position freely in the room can be employed when perceptual aspects of listener movements shall be investigated as e.g. proposed in [34].

5. ACKNOWLEDGEMENTS

The research presented here has been carried out in the Research Project MoNRa which is funded by the Federal Ministry of Education and Research in Germany. Support Code: 03FH005I3-MoNRa. We appreciate the support.
The code of the Matlab-based implementation and a GUI-based version of the *BinRIR* algorithm which is available under the GNU GPL License 4.1 can be accessed via the following webpage:
http://www.audiogroup.web.th-koeln.de/DAFX2017.html

6. REFERENCES

[1] Mel Slater, "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3549–3557, 2009.

[2] Jens Blauert, *Spatial Hearing - Revised Edition: The Psychoacoustics of Human Sound Source Localisation*, MIT Press, Cambridge, MA, 1997.

[3] Etienne Hendrickx, Peter Stitt, Jean-Christophe Messonnier, Jean-Marc Lyzwa, Brian FG Katz, and Catherine de Boishéraud, "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2011–2023, 2017.

[4] W. Owen Brimijoin, Alan W. Boyd, and Michael A. Akeroyd, "The contribution of head movement to the externalization and internalization of sounds," *PLoS ONE*, vol. 8, no. 12, pp. 1–12, 2013.

[5] Ulrich Horbach, Attila Karamustafaoglu, Renato S. Pellegrini, and Philip Mackensen, "Design and Applications of a Data-based Auralization System for Surround Sound," in *Proceedings of 106th AES Convention, Convention Paper 4976*, 1999.

[6] Jens Blauert, Hilmar Lehnert, Jörg Sahrhage, and Holger Strauss, "An Interactive Virtual-Environment Generator for Psychoacoustic Research I: Architecture and Implementation," *Acta Acustica united with Acustica*, pp. 94–102, 2000.

[7] Matthias Geier, Jens Ahrens, and Sascha Spors, "The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods," in *Proceedings of 124th Audio Engineering Society Convention 2008*, 2008, pp. 179–184.

[8] Dirk Schröder and Michael Vorländer, "RAVEN: A Real-Time Framework for the Auralization of Interactive Virtual Environments," *Forum Acusticum*, pp. 1541–1546, 2011.

[9] Juha Merimaa and Ville Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *Journal of the Audio Engineering Society*, vol. 53, no. 12, pp. 1115–1127, 2005.

[10] Ville Pulkki and Juha Merimaa, "Spatial impulse response rendering II: Reproduction of diffuse sound and listening tests," *Journal of the Audio Engineering Society*, vol. 54, no. 1-2, pp. 3–20, 2006.

[11] Ville Pulkki, Mikko-Ville Laitinen, Juha Vilkamo, Jukka Ahonen, Tapio Lokki, and Tapani Pihlajamäki, "Directional audio coding-perception-based reproduction of spatial sound," *International Workshop On The Principles And Applications of Spatial Hearing (IWPASH 2009)*, 2009.

[12] Fritz Menzer, *Binaural Audio Signal Processing Using Interaural Coherence Matching*, Dissertation, École polytechnique fédérale de Lausanne, 2010.

[13] Fritz Menzer, Christof Faller, and Hervé Lissek, "Obtaining binaural room impulse responses from b-format impulse responses using frequency-dependent coherence matching," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 2, pp. 396–405, 2011.

[14] Christoph Pörschmann and Stephan Wiefling, "Perceptual Aspects of Dynamic Binaural Synthesis based on Measured Omnidirectional Room Impulse Responses," in *International Conference on Spatial Audio*, 2015.

[15] Christoph Pörschmann and Stephan Wiefling, "Dynamische Binauralsynthese auf Basis gemessener einkanaliger Raumimpulsantworten," in *Proceedings of the DAGA 2015*, 2015, pp. 1595–1598.

[16] Christoph Pörschmann and Philipp Stade, "Auralizing Listener Position Shifts of Measured Room Impulse Responses," *Proceedings of the DAGA 2016*, pp. 1308–1311, 2016.

[17] Benjamin Bernschütz, "A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100," *Proceedings of the DAGA 2013*, pp. 592–595, 2013.

[18] Benjamin Bernschütz, *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording*, Dissertation, TU Berlin, 2016.

[19] Alexander Lindau, Linda Kosanke, and Stefan Weinzierl, "Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses," *Journal of the Audio Engineering Society*, vol. 60, no. 11, pp. 887–898, 2012.

[20] Philipp Stade, "Perzeptive Untersuchung zur Mixing Time und deren Einfluss auf die Auralisation," *Proceedings of the DAGA 2015*, pp. 1103–1106, 2015.

[21] Philipp Stade and Johannes M. Arend, "Perceptual Evaluation of Synthetic Late Binaural Reverberation Based on a Parametric Model," in *AES Conference on Headphone Technology*, 2016.

[22] Philip Coleman, Andreas Franck, Philip J.B. Jackson, Luca Remaggi, and Frank Melchior, "Object-Based Reverberation for Spatial Audio," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 66–76, 2017.

[23] Wessel Lubberhuizen, "Near perfect reconstruction polyphase filterbank, Matlab Central, www.mathworks.com/matlabcentral/fileexchange/15813 assessed 04/04/2017," 2007.

[24] Benjamin Bernschütz, "MIRO - measured impulse response object: data type description," 2013.

[25] Piotr Majdak, Yukio Iwaya, Thibaut Carpentier, Rozenn Nicol, Matthieu Parmentier, Agnieszka Roginska, Yôiti Suzuki, Kanji Watanabe, Hagen Wierstorf, Harald Ziegelwanger, and Markus Noisternig, "Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions," in *Proceedings of the 134th Audio Engineering Society Convention 2013*, 2013, number May, pp. 262–272.

[26] Benjamin Bernschütz, Christoph Pörschmann, Sascha Spors, and Stefan Weinzierl, "SOFiA - Sound Field Analysis Toolbox," in *Proceedings of the International Conference on Spatial Audio - ICSA*, 2011.

[27] Philipp Stade, Benjamin Bernschütz, and Maximilian Rühl, "A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios," in *27th Tonmeistertagung - VDT International Convention*, 2012, pp. 551–567.

[28] Benjamin Bernschütz, Philipp Stade, and Maximilian Rühl, "Sound Field Analysis in Room Acoustics," *27thth Tonmeistertagung - VDT International Convention*, pp. 568–589, 2012.

[29] Earl G. Williams, *Fourier Acoustics - Sound Radiation and Nearfield Acoustical Holography*, Academic Press, London, UK, 1999.

[30] Philipp Stade, Johannes M Arend, and Christoph Pörschmann, "Perceptual Evaluation of Synthetic Early Binaural Room Impulse Responses Based on a Parametric Model," in *Proceedings of 142nd AES Convention, Convention Paper 9688*, Berlin, Germany, 2017, pp. 1–10.

[31] Fritz Menzer and Christof Faller, "Stereo-to-Binaural Conversion Using Interaural Coherence Matching," in *Proceedings of the 128th AES Convention, London UK*, 2010.

[32] Christoph Pörschmann, Sebastian Schmitter, and Aline Jaritz, "Predictive Auralization of Room Modifications," *Proceedings of the DAGA 2013*, pp. 1653–1656, 2013.

[33] Christoph Pörschmann, Philipp Stade, and Johannes M. Arend, "Binaural auralization of proposed room modifications based on measured omnidirectional room impulse responses," in *Proceedings of the 173rd Meeting of the Acoustical Society of America*, 2017.

[34] Annika Neidhardt and Niklas Knoop, "Binaural walk-through scenarios with actual self-walking using an HTC Vive Real-time rendering of binaural audio Interactive binaural audio scenes Creating scenes allowing self-translation," in *Proceedings of the DAGA 2017*, 2017, pp. 283–286.

4 Auditory Distance Perception of Nearby Sound Sources

4.1 Do Near-Field Cues Enhance the Plausibility of Non-Individual Binaural Rendering in a Dynamic Multimodal Virtual Acoustic Scene?

Arend*, J. M., Ramírez*, M., Liesefeld, H. R., & Pörschmann, C. (2021). *Acta Acust.*, *5*(55), 1–14. (*equal contributions). https://doi.org/10.1051/aacus/2021048

Acta Acustica 2021, **5**, 55

https://doi.org/10.1051/aacus/2021048

Available online at:
https://acta-acustica.edpsciences.org

SCIENTIFIC ARTICLE OPEN ACCESS

Do near-field cues enhance the plausibility of non-individual binaural rendering in a dynamic multimodal virtual acoustic scene?

Johannes M. Arend[1,2,a,*], Melissa Ramírez[1,2,a], Heinrich R. Liesefeld[3], and Christoph Pörschmann[1]

[1] Institute of Communications Engineering, TH Köln – University of Applied Sciences, Betzdorfer Str. 2, 50679 Cologne, Germany
[2] Audio Communication Group, Technical University of Berlin, Einsteinufer 17c, 10587 Berlin, Germany
[3] Department of Psychology, University of Bremen, Hochschulring 18, 28359 Bremen, Germany

Received 27 April 2021, Accepted 8 November 2021

Abstract – It is commonly believed that near-field head-related transfer functions (HRTFs) provide perceptual benefits over far-field HRTFs that enhance the plausibility of binaural rendering of nearby sound sources. However, to the best of our knowledge, no study has systematically investigated whether using near-field HRTFs actually provides a perceptually more plausible virtual acoustic environment. To assess this question, we conducted two experiments in a six-degrees-of-freedom multimodal augmented reality experience where participants had to compare non-individual anechoic binaural renderings based on either synthesized near-field HRTFs or intensity-scaled far-field HRTFs and judge which of the two rendering methods led to a more plausible representation. Participants controlled the virtual sound source position by moving a small handheld loudspeaker along a prescribed trajectory laterally and frontally near the head, which provided visual and proprioceptive cues in addition to the auditory cues. The results of both experiments show no evidence that near-field cues enhance the plausibility of non-individual binaural rendering of nearby anechoic sound sources in a dynamic multimodal virtual acoustic scene as examined in this study. These findings suggest that, at least in terms of plausibility, the additional effort of including near-field cues in binaural rendering may not always be worthwhile for virtual or augmented reality applications.

Keywords: Binaural rendering, Nearby sound sources, Near-field head-related transfer functions, Plausibility, Multimodal environment

1 Introduction

Auditory distance perception is dominated by intensity cues [1, 2]. In reverberant environments, distance judgments are aided by changes in the direct-to-reverberant energy ratio (DRR) [1, 2], and for far-away sources (more than 15 m), high-frequency attenuation provides additional spectral cues [1, 2]. Sound sources in the proximal region[1], i.e., at distances within 1 m of the head center [3], provide further specific distance cues. In particular, interaural level differences (ILDs) exhibit significant distance-dependent changes for lateral sources. Interaural time differences (ITDs), on the other hand, are nearly independent of distance. Both effects were demonstrated by analyses of measured near-field head-related transfer functions (HRTFs) [3, 5]. Brungart [6] suggested that in the absence of the powerful intensity cue, low-frequency ILD cues ($f < 3$ kHz) dominate distance perception of nearby lateral sources in anechoic conditions. Studies by Kopčo et al. on intensity-independent distance perception of nearby sound sources in reverberant conditions yielded inconsistent results, indicating that either the DRR cue masks the ILD cue [7], or that both ILD and DRR cues support distance estimation [8]. Therefore, the relative contribution of the ILD and DRR cues to intensity-independent distance perception is currently not fully understood [9]. Furthermore, nearby sound sources show a relative emphasis of low-frequency sound pressure due to acoustic scattering by the head and torso, resulting in a low-pass filtering character that might be a spectral cue for distance estimation in the near field [1, 3]. The acoustic parallax effect may also affect perception and distance estimation of nearby sound sources [1, 2]. This effect occurs because close sources cause a significant difference between the angle of the source relative to the left or right ear, resulting in a lateral shift of some of the high-frequency features of the HRTF [1].

*Corresponding author: Johannes.Arend@th-koeln.de
[a] Johannes M. Arend and Melissa Ramírez contributed equally to this work.
[1] In the following, we also use the term *near field* to refer to the proximal region [3] or peripersonal space [2], i.e., the area within 1 m of the listener's head center, rather than to describe the frequency-dependent acoustic near field in the sense of physical acoustics [1, 4].

As briefly outlined above, previous research mainly focused on distance estimation accuracy of nearby sound sources and has obtained partly conflicting results regarding the contribution of the various near-field cues to distance perception [1, 2, 6, 10]. A recent study also investigating the influence of binaural cues on distance estimation of nearby sound sources reviews several studies on this topic and discusses the differing results [11]. Further studies in virtual acoustics that used near-field HRTFs synthesized from far-field HRTFs by applying distance variation functions (DVFs) also exclusively evaluated the influence of the synthesized near-field cues on distance estimation accuracy [12, 13]. Moreover, many of the above-mentioned studies tested distance estimation accuracy under unimodal (audio-only) conditions. In most of them, listeners had a passive role (i.e., they could not interact with the sound scene) and had to judge the distance of stationary or dynamic sound events, which, if the study was conducted in virtual acoustics, were even often reproduced with static binaural synthesis only (see, e.g., Arend et al. [11] for an overview). To better evaluate individual auditory distance cues, these methods often attempted to eliminate other cues (e.g., the intensity cue by level normalization), resulting in unnatural stimuli. Thus, whereas such experimental methods are well suited to understand the contribution of individual distance cues to distance perception and how they interact with each other, they do not ideally reflect the way humans perceive their multimodal environment and estimate, for example, the distance to a (nearby) sound source in real-life.

Rummukainen et al. [14] presented the only study we are aware of that investigated perceptual aspects of near-field HRTFs beyond distance estimation accuracy, and that was conducted in a six-degrees-of-freedom (6-DoF) multimodal virtual reality (VR) environment, thereby including visual and proprioceptive cues in addition to auditory cues. In their experiment, listeners either actively moved around a static virtual sound source or dynamically moved the virtual sound source around their head. The participants' task was to rate binaural renderings based on intensity-scaled far-field HRTFs or multi-distance near-field HRTFs, among others, according to their preference. Surprisingly, listeners liked both HRTF types equally. However, the authors pointed out that further studies are needed, especially as the closest distance examined in their study was 0.50 m, which means that the strongest near-field cues were not present.

Thus, whereas it is generally assumed that including near-field cues in binaural rendering leads to a more realistic reproduction, and especially experienced listeners often report that near-field effects are subjectively audible, studies such as Rummukainen et al. [14] raised first doubts on the perceptual importance of near-field HRTFs in multimodal environments. However, to the best of our knowledge, no study has examined yet whether using near-field HRTFs for binaural rendering in a dynamic multimodal scene enhances the *plausibility* [15] of the virtual acoustic environment (VAE) compared to using intensity-scaled far-field HRTFs, i.e., whether using near-field HRTFs results in a binaural reproduction of nearby sound sources that, based on the listener's *inner reference* and personal experience, is more in agreement with their expectation towards the corresponding real event than binaural rendering using intensity-scaled far-field HRTFs.

The plausibility of virtual environments has been discussed extensively in the literature of various research areas, and the above-mentioned definition by Lindau & Weinzierl [15] is in line with what Slater [16] referred to as *plausibility illusion* and Hofer et al. [17] recently described as *external plausibility*. Essentially, external plausibility refers to how *consistent* the virtual environment is with the users' real-world knowledge [17], and whether an event in the virtual environment could actually occur in the real world [16]. Thus, external plausibility is expressed by the user (or more precisely, in this case, the listener) judging something in the virtual environment to be factually true or accurate, or by events in the virtual environment to be highly likely or typical of the real world [17].

Assessing the plausibility of a VAE provides, therefore, a comprehensive measure for the quality of the virtual presentation that includes various perceptual factors. It is an important perceptual criterion for VR and augmented reality (AR) applications, as its assessment also examines how the acoustic representation agrees with other modalities of the virtual scene (e.g., visual, haptic, or proprioceptive) and whether there are no apparent contradictions between the modalities that would reduce or even break plausibility [18]. As such, plausibility has recently become a popular measure for the perceptual evaluation of VR and AR audio applications.

However, the various audio and acoustic studies that have assessed the plausibility of VAEs often differ in their experimental methods and procedures. Lindau & Weinzierl [15] proposed a test paradigm in which either a real (loudspeaker reproduction) or a virtual (binaural reproduction) stimulus is presented in each trial, and the participants have to decide in a yes/no task whether the stimulus comes from a real loudspeaker or a virtual representation of the loudspeaker. Some studies followed this procedure, in which the stimulus is presented either through a real loudspeaker or binaurally through headphones, for example, to evaluate the plausibility of pseudobinaural recordings [19], or 6-DoF parametric binaural rendering [20]. However, the test paradigm proposed by Lindau & Weinzierl [15] has also been adapted (by the same research group) to assess the plausibility of room acoustic simulations. In the study by Brinkmann et al. [21], there was no real source serving as an explicit reference. Instead, listeners were presented with either simulation- or measurement-based auralizations (only virtual stimuli) and had to rate whether the stimuli correspond to a real room. In line with this, several other approaches have been proposed to assess the plausibility of VAEs in cases where no real counterpart is available to use as an explicit reference. For example, Neidhardt et al. [22, 23] evaluated the plausibility of position-dynamic virtual acoustic realities in which listeners move towards a virtual sound source using either a continuous or ordinal plausibility rating scale. Amengual Garí et al. [24] evaluated

the plausibility of 3-DoF parametric binaural rendering using a two-alternative forced-choice (2AFC) procedure. Either both stimuli were virtual, or one of them was a real loudspeaker, and participants had to rate which of the two stimuli they perceived as more plausible. Most recently, Neidhardt & Zerlik [25] conducted two experiments to assess the plausibility of position-dynamic binaural rendering using a yes/no task. In one experiment, participants were presented with virtual stimuli only, whereas in another experiment, participants were presented with either real or virtual stimuli. The authors concluded that because of their different advantages and disadvantages, both methods are relevant and valid for assessing the plausibility of a VAE.

Surprisingly, even though very recent research such as VRACE [26] focuses on binaural rendering in the near field, and although several binaural renderers use near-field cues or respectively (synthesized) near-field HRTFs to reproduce nearby sound sources (e.g., the commercially available renderers from Oculus [27], MagicLeap [28], and Resonance Audio [29] as well as the open-source renderers Spat [30], Anaglyph [31], or 3DTI Toolkit [32]), it is still unknown whether binaural rendering with near-field HRTFs increases the plausibility for naive (non-expert) listeners compared to a much easier to implement rendering with intensity-scaled far-field HRTFs. However, it is crucial to know whether the additional computing effort of including near-field cues is worthwhile in terms of plausibility and overall reproduction quality, especially for complex real-time applications with limited computing resources, such as mobile AR applications with 6-DoF.

To close this gap and investigate whether near-field HRTFs provide a more plausible binaural reproduction of nearby sound sources than intensity-scaled far-field HRTFs, we performed two listening experiments in an anechoic 6-DoF VAE. In both experiments, participants controlled the position of a virtual sound source by moving a small handheld loudspeaker, which provided visual and proprioceptive cues in addition to the auditory cues and aided the application-oriented AR experience. In a 2AFC procedure, the participants had to compare non-individual anechoic binaural renderings based on either synthesized near-field HRTFs or intensity-scaled far-field HRTFs and judge which of the two rendering methods led to a more plausible representation, i.e., which one was more congruent with their expectations based on the visual- and haptic sensation as well as based on their inner reference and personal experience. We hypothesized that they would rate the renderings using near-field HRTFs as more plausible, as this reproduction method yields a more physically correct representation of nearby sound sources.

We employed a multimodal sensory-motor test paradigm where participants moved the sound source because this results in a more natural scenario that better emulates the way humans perceive their environment than the more extensively investigated unimodal passive paradigms. Besides, previous research showed that multisensory stimulation improves sound localization [33, 34]. Recent findings by Valzolgher et al. [35] also indicated that kinesthetic cues resulting from moving a sound source with one's own hand could contribute to the updating of spatial hearing and thus improve sound localization performance. In this line of thinking, providing more reliable (real) visual, motor, and proprioceptive information simultaneously, together with the (simulated) auditory information, should help listeners optimally associate auditory cues to the spatial location of a sound source [34, 35]. Thus, the multimodal virtual environment employed in our study should (1) facilitate auditory localization and (2) provide the listener with more information to assess the plausibility of a virtual sound source more reliably than is possible in a unimodal environment. Listeners were able to judge the plausibility of the binaural renderings based not only on their inner reference and listening experience but also on the simultaneous real information (visual, motor, and proprioceptive). This aided the identification of possible discrepancies between the real and virtual worlds and thus detecting breaks in plausibility.

The two experiments, each performed with a different group of subjects, differed only regarding the test signal used. In Experiment 1, we used pink noise bursts to provide extremely critical and ideally controllable stimuli that clearly reveal all near-field cues. Then, to generalize the results of Experiment 1 to a more application-oriented setup, we used female speech as a test signal in Experiment 2.

2 Experiment 1

2.1 Method

2.1.1 Participants

Sixteen participants (ages 22–55 years, $M = 33.3$ years, $Mdn = 28$ years, $SD = 11.2$) with self-reported normal hearing took part in the experiment on a voluntary basis. Four of the participants are members of our laboratory and therefore classified as expert listeners. The remaining participants were engineering students or research assistants from other laboratories at the university and classified as naive listeners. All participants were naive as to the purpose of the study.

2.1.2 Setup

The experiment took place in the sound-insulated anechoic chamber of TH Köln, which provided the appropriate acoustic environment for the anechoic binaural renderings simulating the handheld loudspeaker. The experiment was implemented, controlled, and executed by a purpose-built Python application running on a PC. For real-time dynamic binaural synthesis, we employed the open-source tool PyBinSim [36] in combination with a pair of HTC VIVE trackers (update rate of 120 Hz). One tracker was mounted on the headphones (Sennheiser HD600), and the other tracker was attached to the handheld loudspeaker (JBL Clip+), providing 6-DoF tracking data of both. Based on the tracking data, the Python application calculated the loudspeaker's azimuth, elevation, and distance relative to

the participant's head orientation and position and sent these spherical coordinates to PyBinSim by Open Sound Control (OSC) messages. The application also used OSC messages to control the renderer, e.g., to start and stop audio playback or to change between HRTF datasets. Additionally, the application logged the relative tracking data at a sampling rate of 30 Hz.

The graphical user interface of the application was presented on a screen located at a distance of about 2 m in front of the seated participant. A Numark Orbit MIDI controller served as the input device for the participants' responses. We used an RME Babyface audio interface as digital-to-analog converter and headphone amplifier at 48 kHz sampling rate and a buffer size of 64 samples. The separate buffer of PyBinSim was set to 128 samples.

2.1.3 Materials

We employed measured far-field HRTFs from a Neumann KU100 dummy head [37], a dataset widely used in both commercial applications and research. The HRTF set was transformed to the spherical harmonics (SH) domain at a sufficiently high spatial order of $N = 44$, allowing artifact-free SH interpolation to obtain HRTFs for any desired direction, which was necessary in the present case for accurate HRTF synthesis. Both the intensity-scaled far-field HRTFs as well as the near-field HRTFs were synthesized for distances from 0.12 m to 1.20 m in steps of 1 cm on a spatial sampling grid with a resolution of 1° in the horizontal direction and 5° in the vertical direction, limited to ±15° in elevation.

The near-field HRTFs were synthesized by applying distance variation functions (DVFs) to the far-field HRTFs [12]. The DVFs were generated from a spherical head model [38] with the ears positioned at azimuth $\phi = \pm 90°$ and elevation $\theta = 0°$. The optimal head radius of the spherical head model was 9.19 cm, calculated according to Algazi et al. [39] based on the dimensions of the Neumann KU100 dummy head. In general, DVFs are calculated for each distance and direction as the ratio of the pressure on the sphere emanating from a sound source at a desired distance in the near field to the pressure on the sphere emanating from a sound source in the far field, with the pressure on the sphere evaluated solely at the ear positions. Thus, a DVF approximates the changes of an HRTF as a sound source varies in distance, such as alterations in intensity and spectrum or frequency-dependent changes in ILD. Additionally, a cross-ear parallax correction was applied [40] to account for high-frequency parallax effects induced by the pinna, which the DVF is unable to take into account [12]. Appropriate far-field HRTFs for the left and right ear are first selected for the respective distance and direction (using SH interpolation) based on a geometric parallax model and then filtered with the corresponding DVFs, resulting in the desired near-field HRTFs. The described processing, which is similar to the implementation in state-of-the-art renderers such as Spat, Anaglyph, or 3DTI Toolkit, was performed using the `supdeq_dvf` function of the `SUpDEq toolbox`[2].

To synthesize the intensity-scaled far-field HRTFs, an HRTF set was first obtained by SH interpolation according to the spatial sampling grid, and then its level was matched to that of the near-field HRTF set for the highest distance of 1.20 m. This set was then adjusted in level according to the inverse-square law to generate the HRTFs for closer distances. Thus, the intensity-scaled far-field HRTFs do not contain any of the prominent near-field cues included in the synthesized near-field HRTFs, such as the significant increase in (low-frequency) ILD for lateral sources, the low-pass filtering character, and the parallax effects.

Figure 1 (left) shows the low-frequency ($f < 3$ kHz) horizontal plane ILDs (which Brungart [6] suggests are the dominant auditory distance cue in the near field) for the intensity-scaled far-field HRTF sets (FF) and synthesized near-field HRTF sets (NF) at selected distances. As expected, the ILDs of the near-field HRTFs for lateral sources increase strongly with decreasing distance, especially for close distances (less than 0.50 m). The right plot in Figure 1 shows the corresponding ITDs, which, as expected, are nearly distance-independent and therefore almost the same for all HRTF sets. Figure 2 further shows the frequency-dependent behavior of the ILDs of the synthesized near-field HRTF sets as a function of distance. Consistently, the synthesized near-field HRTFs show strong low-frequency ILDs for lateral directions at close distances and a significant increase in ILD with increasing frequency. Overall, the described characteristics of the synthesized HRTFs are very similar to those of measured near-field HRTFs [5, 11, 41], confirming that the synthesis yields correct results. In particular, the low-frequency horizontal plane ILDs and ITDs of the synthesized near-field HRTFs are nearly identical to those of measured Neumann KU100 near-field HRTFs from [5] (see Fig. S1 in the Supplementary Material [42]), further supporting the excellent performance of the synthesis.

The test signal was a 10 s long sequence of 500 ms pink noise burst (including 10 ms cosine-squared onset/offset ramps) with an interstimulus interval of 150 ms. Broadband noise bursts are well-suited test signals to examine coloration and localization, so they were ideal for the present experiment. The sequence length of 10 s provided sufficient time to move the loudspeaker along the prescribed trajectory (see procedure in Sect. 2.1.4). To minimize the influence of the Sennheiser HD600 headphones, a generic headphone compensation filter was used. The filter was based on 12 measurements in which the headphones were put on and off the Neumann KU100 dummy head (the same one used to measure the far-field HRTFs employed in the present study) to account for re-positioning variability. The final filter was designed by regularized inversion of the complex mean of the headphone transfer functions [43] using the implementation by Erbes et al. [44]. Furthermore, to enhance the virtual acoustic representation of the handheld JBL Clip+ loudspeaker, a filter describing its on-axis frequency response was designed. The magnitude

[2] Available: https://www.github.com/AudioGroupCologne/SUpDEq

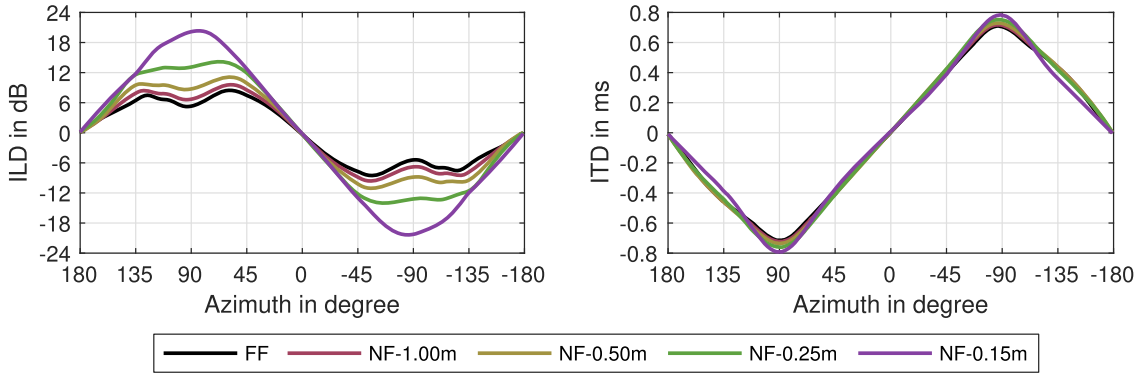


Figure 1. Low-frequency ($f < 3$ kHz) horizontal plane ILDs (left) and ITDs (right) of the intensity-scaled far-field HRTF sets (FF) and synthesized near-field HRTF sets (NF) at selected distances.

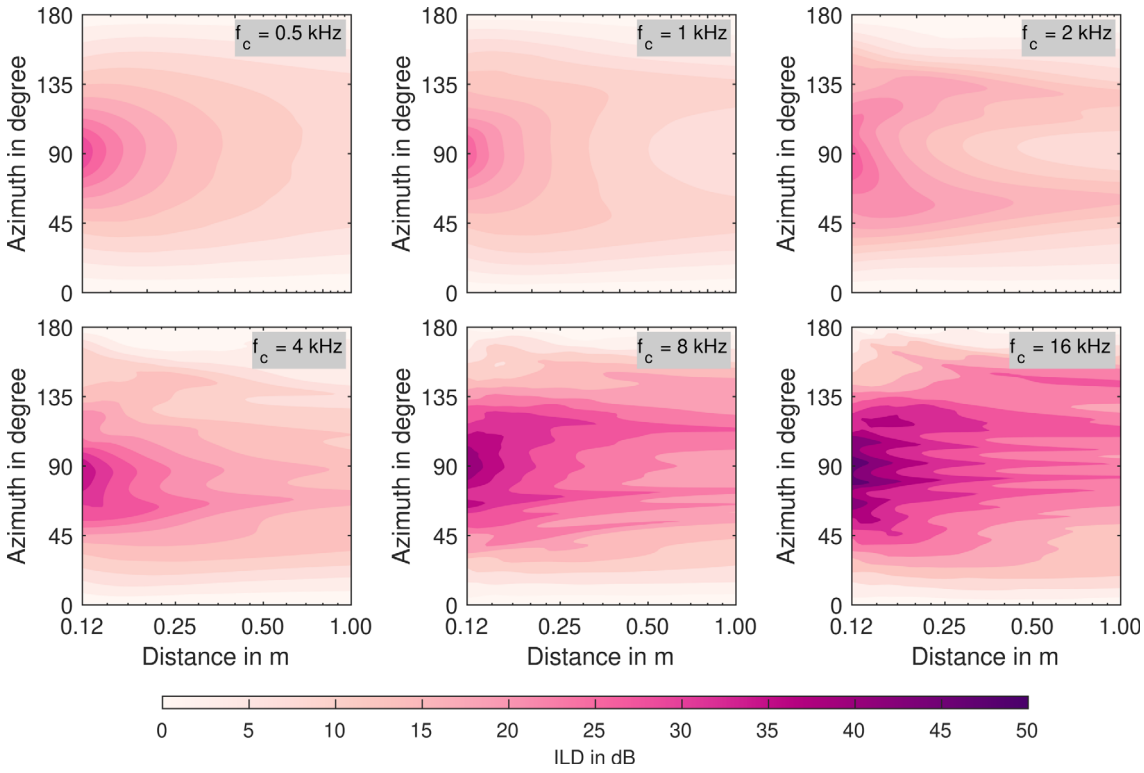


Figure 2. Horizontal plane ILDs (left hemifield) of the synthesized near-field HRTF sets as a function of distance for octave bands from 0.5 kHz to 16 kHz. Distances are shown on a logarithmic scale for a more detailed representation of the ILDs at close distances.

responses of both filters were combined to one minimum-phase finite impulse response (FIR) filter with 2048 taps, which was applied to the test signal. For more technical details, Figure S2 in the Supplementary Material [42] shows the magnitude response of the employed headphone compensation and loudspeaker filter. Informal evaluations showed that the 200 Hz low-cut of the loudspeaker filter does not affect the (binaural) near-field cues of the synthesized HRTFs. However, in pilot studies, we found that applying the filter is essential for the plausibility of the multimodal scene. Filtering out the low frequencies of the stimuli aligns the auditory impression with the visual impression of a small handheld loudspeaker. Besides, to foster reproducible research, we provide as well in the Supplementary Material [42] the Matlab script developed to synthesize the near- and far-field HRTFs, design the filters, and generate the filtered test signal.

To measure the presentation level produced over the headphones, a loudspeaker in the free field was leveled so that the playback of stimuli for frontal sound incidence produced the same electrical level at a dummy head as their playback over the headphones on the dummy head. The presentation level was then measured as the loudspeaker's equivalent free-field sound pressure level directly at the

dummy head's ear. Following this procedure, we estimated the presentation level for different conditions (without roving, meaning at a roving level of 0 dB; see roving procedure described in Sect. 2.1.4). The measured presentation level of the far-field condition for frontal sound incidence was $L_{\mathrm{Aeq}} = 49.3$ dB for a distance of 1.00 m and $L_{\mathrm{Aeq}} = 69.6$ dB for a distance of 0.12 m. The highest presentation level was $L_{\mathrm{Aeq}} = 84.5$ dB, measured for lateral sound incidence at the closest distance (0.12 m) in the near-field condition.

2.1.4 Procedure

Participants directly compared dynamic binaural renderings based on the intensity-scaled far-field HRTFs with renderings based on the synthesized near-field HRTFs in a 2AFC procedure. Each of the 100 trials in total consisted of a sequence of two 10 s intervals with an interstimulus interval of 0.5 s. The presentation order, i.e., whether the far-field or near-field rendering was presented first, was randomized. Moreover, the presentation level of each interval was randomly roved within a 10 dB range (±5 dB, steps of 1 dB, see, e.g., Kopčo & Shinn-Cunningham [7]) and participants were informed about that.

During the presentation of each interval, the participants were asked to move the handheld loudspeaker along a prescribed square-like trajectory to direct the virtual sound source through frontal and lateral areas near the head that yield strong near-field cues and thus clear differences between the rendering conditions. As a result, participants were exposed to all relevant auditory near-field cues: (1) frequent distance changes of the virtual sound source in lateral areas yielded strong variations in (low-frequency) ILD cues and distinct intensity cues, (2) movements of the virtual sound source from lateral to frontal areas very close to the head provided significant spectral, ILD, parallax, and intensity cues, and (3) frequent distance changes of the virtual sound source in frontal areas yielded strong spectral, parallax, and intensity cues.

After the presentation of both intervals, they were asked to select the interval which, as verbally instructed before the experiment, provided a more accurate representation of the expected sound field according to the sound source's positions and movements. In other words, participants had to choose the more plausible sound field representation based on their inner reference [15], life experience, and auditory, proprioceptive, and visual cues that emerged from actively moving the virtual source. The participants gave their answer by pressing a button on the MIDI controller. The answer was scored as correct when participants chose the near-field condition, following our initial hypothesis that using near-field HRTFs should be perceived as more plausible because it yields a more physically correct representation of nearby sound sources. Participants could neither repeat a trial nor continue without answering, and no feedback was provided. After an answer was registered, there was a 1 s silent pause before the next trial started. The procedure, including a presentation of the prescribed trajectory, is also illustrated in a short video, which is part of the Supplementary Material [42].

The 100 trials were split into two blocks of 50 trials with a short break in between to prevent fatigue. Before the experiment, participants were given instructions about the experimental procedure and they had to perform two training blocks to get familiar with the setup and the test procedure. In the first training block, participants were asked to practice moving the handheld loudspeaker along the prescribed trajectory. Their actual movement trajectory was displayed in real time on a computer screen so that they could visually monitor whether it conformed with the prescribed trajectory and adapt the movement trajectory based on this feedback if necessary. In the second training block, participants had to perform five trials of the experiment to practice the test procedure while still receiving the on-screen feedback. After the training, participants had no on-screen feedback on their movements to not distract them from the main task. A complete experimental session lasted about one hour, including the verbal instructions, the training blocks, and the short break.

2.2 Results and discussion

In informal post-experiment interviews, participants were asked whether the binaural reproduction was generally plausible regardless of the rendering condition. Overall, they experienced the scene as plausible, i.e., they perceived that the real loudspeaker emitted the sound, and they localized the virtual source at the position of the real loudspeaker. They reported that, in particular, moving the source and the congruence of visual, proprioceptive, and auditory cues supported the plausibility of the scene.

To verify that participants moved the loudspeaker mainly along the prescribed trajectory, we first analyzed the movement patterns based on the tracking logs. Figure 3 (left) shows the relative tracking data of Experiment 1, pooled over all participants and trials, in the form of a two-dimensional histogram. The plot shows the frequency distribution of the sound source position relative to the participants' head in the horizontal plane, defined by azimuth and distance. The prominent square-like movement pattern reflects the prescribed trajectory. As instructed, participants varied the distance of the virtual sound source to a large extent at frontal and lateral azimuth angles, which resulted in significant spectral (frontal) and ILD (lateral) changes in the near-field condition and intensity changes in both conditions. Furthermore, participants often placed the virtual source very close, both frontally and laterally, at distances between 0.20 m and 0.30 m. This also provided strong spectral (frontal) and ILD (lateral) cues in the near-field condition and thus significant differences to the far-field condition, at least from a signal-theoretic point of view. For a more detailed analysis, we provide plots of each participant's individual movement pattern in Figure S3 of the Supplementary Material [42].

Figure 4 (left) shows the results of the experiment in terms of individual p_{2AFC} values, their mean, and their 95% between-subject confidence interval (CI). The right plot of Figure 4 shows the interindividual variation in the determined p_{2AFC} values in the form of a box plot.

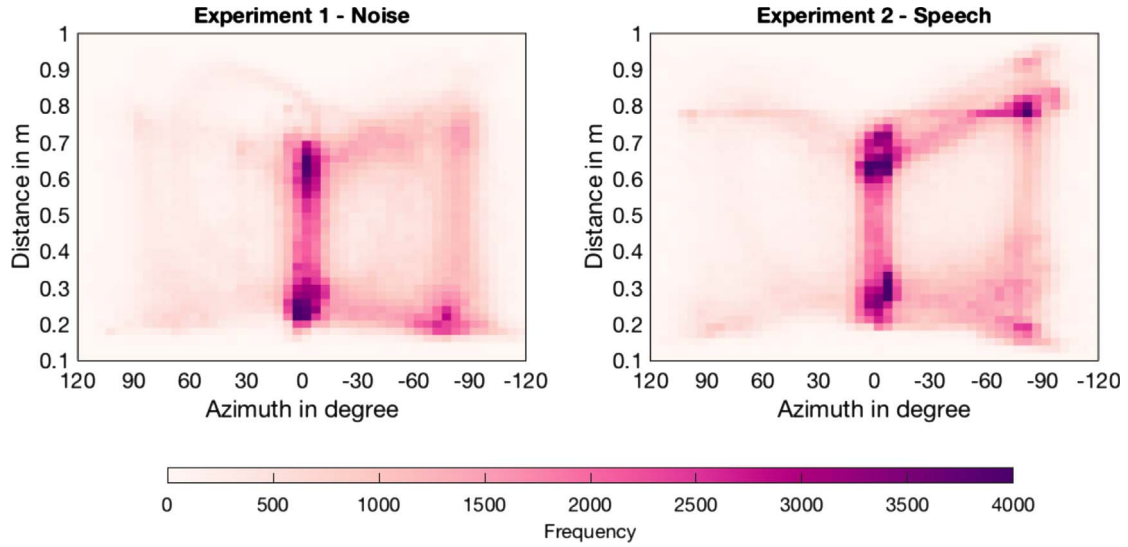


Figure 3. Two-dimensional histograms of the relative tracking data of Experiment 1 (left) and Experiment 2 (right), pooled over all participants and trials in the respective experiment.

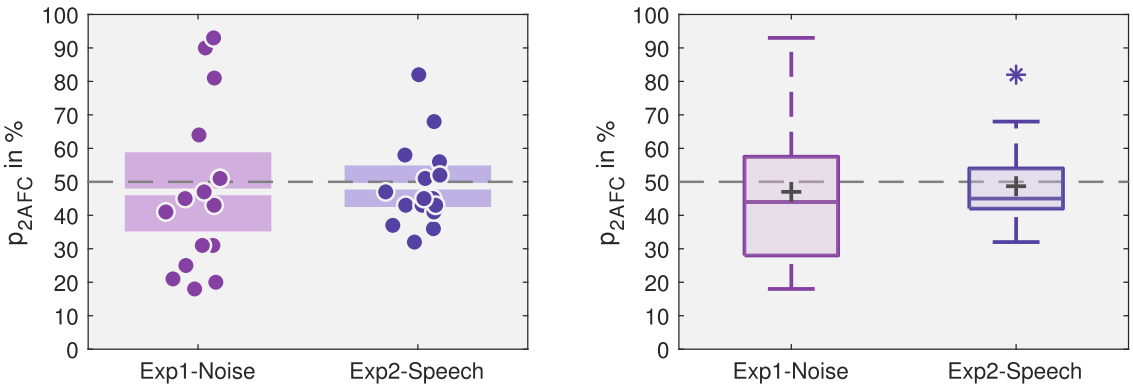


Figure 4. Results of the 2AFC test in Experiment 1 (Exp1-Noise) and Experiment 2 (Exp2-Speech). The left plot shows the determined individual percentages of correct answers p_{2AFC} as points (horizontal offset for better readability). The boxes show the mean (box notch) and the 95% between-subject CI. The gray dashed line denotes 50% chance level. The right plot shows the interindividual variation in the determined p_{2AFC} values in the form of a box plot with the median (box line), the mean (cross), and the (across participants) interquartile range (IQR); whiskers display 1.5 × IQR below the 25th or above the 75th percentile and outliers beyond that range are indicated by asterisks.

In general, the results exhibit high between-subject variance (see left plot in Fig. 4). Two participants, which both are expert listeners, performed exceptionally well (p_{2AFC} = 90% and 93%), but the majority of the participants either performed near 50% chance level or even clearly below chance. The findings suggest that the two participants strongly favored the near-field condition, whereas most other participants could not decide which condition was a more plausible reproduction (near chance performance), or even preferred the far-field condition over the near-field condition (below chance performance). Consequently, the mean and the median are slightly below chance level (see right plot in Fig. 4).

For statistical analysis of the results, we first applied a Lilliefors test for normality to the p_{2AFC} values, which showed no violations of normality (p = .151), indicating that parametric tests can be used. To analyze if the p_{2AFC} mean differs significantly from chance, we performed a one-sample t test against 50%. The test yielded no significant difference between the p_{2AFC} mean of 47% and chance level [$t(15)$ = 0.50, p = .626, d = .12]. As non-significant results of null-hypothesis significance testing cannot be interpreted as evidence for the absence of an effect, we also calculated the respective Bayes factor (BF_{01}, JZS scaling factor r = .707) for the one-sample t test. The obtained BF_{01} = 3.51 suggests that the data provide more than 3 times more evidence for the absence (rather than the presence) of an effect of near-field cues. Thus, the statistical results confirm that, on average, participants could not reliably decide which rendering method was more plausible, or in other words, on average, they found both rendering methods equally plausible.

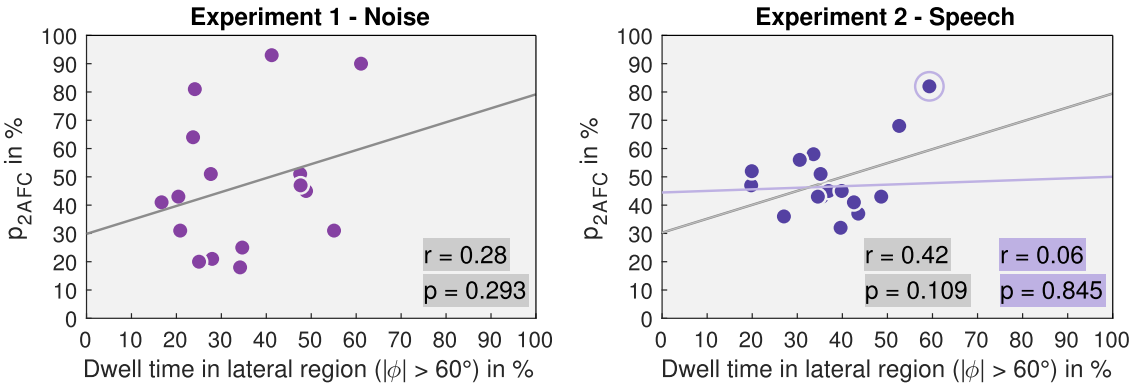


Figure 5. Participants' dwell time in lateral region ($|\phi| > 60°$) vs. their percentage of correct answers p_{2AFC} for Experiment 1 (left) and Experiment 2 (right). The solid line is the least-squares line of best fit. The right plot shows in purple the results of the correlation analysis for Experiment 2, excluding the outlier (data point circled in purple).

Next, we analyzed whether there is a correlation between participants' movement patterns and their plausibility estimates. For lateral source positions, the near-field HRTFs additionally exhibit strong ILD cues, resulting in particularly severe differences between the near- and far-field conditions. If these ILD cues affect listeners' preferences, there might be a correlation between the time subjects spend in lateral regions (*dwell time* in the following) and the percentages of correct answers. In other words, we examined whether participants who more often positioned the virtual source laterally perceived the near-field condition as more plausible. For this, we calculated the Pearson correlation between the participants' dwell time in the lateral region (proportion of relative tracking data with $|\phi| > 60°$, according to the definition of lateral positions by Brungart [6]) and the p_{2AFC} values. This yielded a non-significant positive correlation between dwell time and the p_{2AFC} values [$r(14) = .28$, $p = .293$], providing no evidence that participants who frequently positioned the virtual source laterally chose the near-field condition more often as the most plausible. Figure 5 (left) shows the corresponding scatter plot illustrating the relationship between both variables.

Finally, to determine whether plausibility ratings changed over the course of the experiment, e.g., because participants became tired or learned certain stimuli features, we analyzed the p_{2AFC} values in four epochs of 25 trials each. Figure 6 (left) shows the results of the experiment, divided among the four epochs. The plots suggest that participants remained fairly consistent in their answers over time. Thus, most participants who perceived the near-field condition as more plausible at the beginning of the experiment also did so throughout the experiment. The response behavior is similarly consistent for participants who preferred the far-field condition or perceived both conditions as equally plausible. In general, the between-subject variance seems to increase slightly over time, as participants who preferred the near- or far-field condition in particular became more stringent (more extreme) throughout the experiment, tending toward $p_{2AFC} = 100\%$ and $p_{2AFC} = 0\%$, respectively.

Statistical analysis of the data concerning the factor epoch showed no significant effect, suggesting that neither learning nor fatigue effects had a systematic impact on the participants' average responses. In particular, Greenhouse-Geisser (GG) corrected [45] one-way repeated measures ANOVA with the within-subject factor epoch revealed no significant effect of epoch [$F(3,45) = 1.38$, $p = .261$, $\eta_p^2 = .08$, $\epsilon = .62$]. In line with this, a paired t test comparing the results of the first and fourth epoch yielded no significant difference [$t(15) = 0.47$, $p = .646$, $d_z = .12$], indicating that participants answered similarly at the beginning and end of the experiment. The respective Bayes factor analysis for this pairwise comparison provided some evidence for the absence of an effect of epoch ($BF_{01} = 3.55$). Finally, a Levene's test comparing the between-subject variance of the results in the first and fourth epoch yielded no significant difference [$F(1,30) = 1.66$, $p = .208$], thus providing no statistical support for the observations that participants' answers might become more extreme towards the end of the experiment.

To quantify how consistent participants preferred one over the other rendering method across the experiment, we calculated Pearson correlations between all pairs of epochs. As shown in Table 1, these correlations were high and significant throughout, demonstrating that participants' preferences were highly consistent across epochs. By implication, the high correlations additionally show that at least those participants who strongly favored the near- or far-field condition were able to clearly discriminate the respective HRTFs.

In addition, we also examined participants' individual movement patterns across epochs (see Figs. S5–S8 in the Supplementary Material [42]). The plots show that most participants consistently performed similar movements and did not notably change their movement pattern during the experiment. These observations may indicate that, as we expected, the movement actually became automatic for participants after a short period of time (already during training or within the first few trials of the first epoch), allowing them to focus their cognitive resources on the

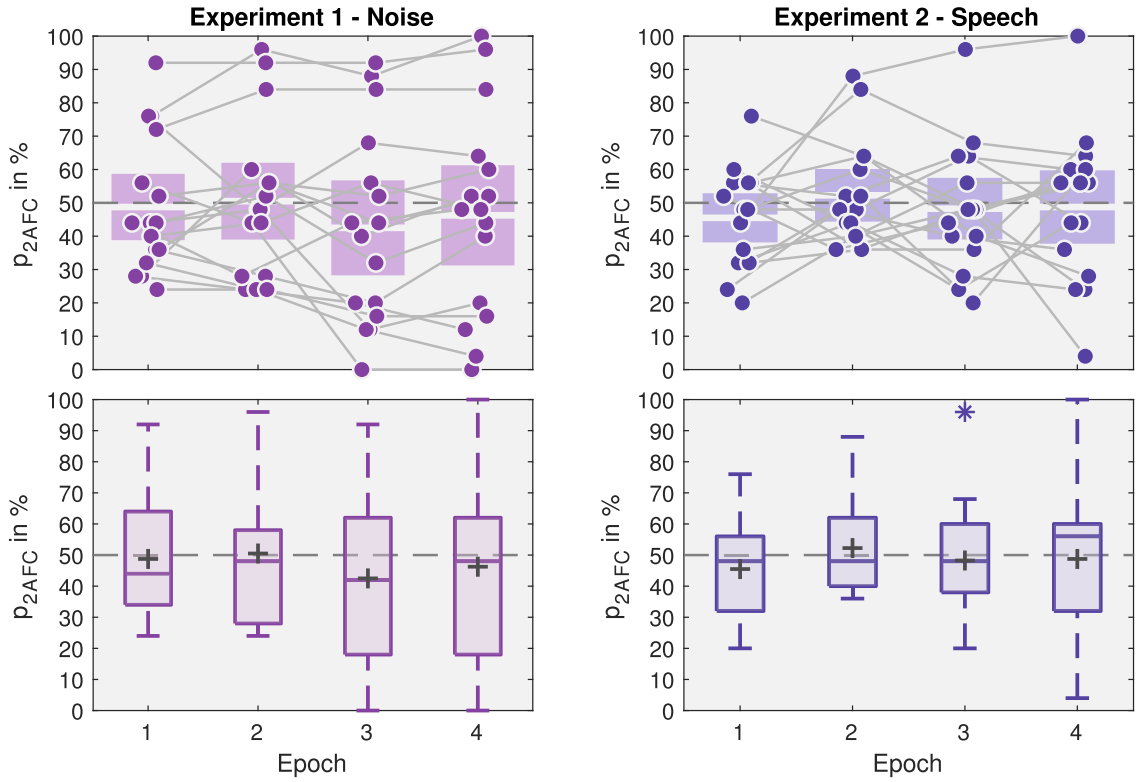


Figure 6. Results of Experiment 1 (left) and Experiment 2 (right) as a function of four epochs with 25 trials each. The top plots show the determined individual p_{2AFC} values in each epoch as points (horizontal offset for better readability). Individual data points are connected with gray lines. The boxes show the mean (box notch) and the 95% between-subject CI for each epoch. The gray dashed line denotes 50% chance level. The bottom plots show the interindividual variation in the determined p_{2AFC} values for each epoch in the form of a box plot with the median (box line), the mean (cross), and the (across participants) interquartile range (IQR); whiskers display 1.5 × IQR below the 25th or above the 75th percentile and outliers beyond that range are indicated by asterisks.

Table 1. Pearson correlation coefficients between all pairs of epoch for Experiment 1 (top) and Experiment 2 (bottom).

Epoch	1	2	3	4
	Experiment 1 – Noise			
1				
2	.78***			
3	.75***	.81***		
4	.72**	.85***	.96***	
	Experiment 2 – Speech			
1				
2	.34			
3	.12	.60*		
4	−.19	.44	.51*	

Note. $^*p < .05$, $^{**}p < .01$, $^{***}p < .001$, $N = 16$.

listening task rather than on moving the handheld loudspeaker (see, e.g., [46, 47]).

3 Experiment 2

The results of Experiment 1 provided no evidence that near-field cues enhance the plausibility of binaural rendering in a dynamic multimodal virtual acoustic scene as employed in this study. One possible explanation for these rather surprising results is that the pink noise test signal used in Experiment 1 is perceived as unnatural no matter the plausibility of the HRTFs, because pink noise rarely occurs in everyday situations. Thus, participants have no listening experience with such a stimulus and therefore might find it difficult to judge its plausibility based on their life experience and inner reference. For this reason and to provide a stimulus more commonly encountered in the near field, we used female speech as the test signal in Experiment 2, which was otherwise identical in design and procedure to Experiment 1. Furthermore, using a speech stimulus makes Experiment 2 more similar to applied scenarios, as near-field rendering of speech is important for various VR and AR applications.

3.1 Method

A new sample of 16 participants (ages 20–33 years, $M =$ 25.8 years, $Mdn = 28$ years, $SD = 3.9$) with self-reported normal hearing took part in the experiment for course credit. All participants were engineering students without experience in listening experiments and therefore classified as naive listeners. They were all naive as to the purpose of the study.

As outlined above, the only difference from the first experiment was that we used female speech as the test signal in this experiment. We chose the first, second, third, and sixth phonetically balanced sentences from the first list of Harvard sentences, spoken by a native female British English speaker [48]. The sentences were composed into a sequence of 10 s length (the same length as the noise burst sequence used in Experiment 1) with 62.5 ms silent pauses between the sentences. Similar to the first experiment, the speech test signal was filtered with the minimum phase FIR filter, combining the headphone compensation filter and the loudspeaker filter. To ensure similar presentation levels as in Experiment 1, the loudness of the speech test signal was adjusted to that of the noise test signal used in Experiment 1 according to the ITU-R BS.1770-4 recommendation [49]. The described processing can be reproduced by the Matlab script available in the Supplementary Materials [42]. In all other aspects, setup, materials, procedure, and analysis were identical to Experiment 1 (see Sect. 2).

3.2 Results and discussion

Participants in Experiment 2 also generally perceived the scene as plausible, as determined by informal post-experiment interviews. Figure 3 (right) shows the two-dimensional histogram of the relative tracking data of Experiment 2, pooled over all participants and trials. Again, it shows the square-like movement pattern reflecting the prescribed trajectory. Thus, participants in Experiment 2 also very frequently covered positions that yielded strong near-field cues in the near-field condition. Figure S4 in the Supplementary Material [42] provides individual-subject data.

Figure 4 also shows the results of Experiment 2. The majority of the participants performed near chance level (see left plot in Fig. 4), indicating that most participants could not decide which rendering method was more plausible or simply perceived both conditions as equally plausible. Only a single (outlying) participant (see right plot in Fig. 4) clearly perceived the near-field condition as more plausible than the far-field condition. Consequently, the box plot in Figure 4 (right) exhibits a rather small IQR with the mean and median slightly below chance level.

A Lilliefors test for normality showed no violations of normality ($p = .178$), so we performed a one-sample t test against chance level. In line with the plots, the test yielded no significant difference between the p_{2AFC} mean of 48.7% and chance level [$t(15) = 0.41$, $p = .684$, $d = .10$]. The respective Bayes factor analysis provided some evidence for the absence of the effect ($BF_{01} = 3.63$).

Participants' dwell time in the lateral region ($|\phi| > 60°$) did not significantly correlate with their performance [$r(14) = .42$, $p = .109$]. A close look at the corresponding scatter plot in Figure 5 (right) indicates that the (sizeable) correlation is mainly driven by the outlier, dropping to $r(13) = .06$, $p = .845$ with this outlier excluded (see results in purple in the right plot of Fig. 5). Thus, for the vast majority of participants, there is no evidence that they perceived the near-field condition as more plausible even when they frequently positioned the virtual sound source in lateral regions, producing strong binaural near-field cues and clear differences between near- and far-field conditions.

Analysis of plausibility ratings in the four epochs (each with 25 trials) showed that the majority of participants consistently performed close to chance throughout the experiment (see Fig. 6 (right)). Only one participant (the outlier) clearly tended increasingly towards the near-field condition over epochs. Thus, considering the entire data set, we did not detect a significant fatigue or learning effect. A GG-corrected one-way repeated measures ANOVA with the within-subject factor epoch showed no significant effect of epoch [$F(3,45) = 0.52$, $p = .668$, $\eta_p^2 = .03$, $\epsilon = .72$], and a paired t test comparing the results of the first and fourth epoch also showed no significant difference [$t(15) = 0.44$, $p = .664$, $d_z = .38$]. The Bayes factor analysis for the latter pairwise comparison yielded some evidence for the absence of an effect of epoch ($BF_{01} = 3.59$). Comparing the variances of the results in the first and fourth epoch with a Levene's test again yielded no significant difference [$F(1,30) = 1.48$, $p = .234$], thus providing no indication that answers would become more extreme across the course of the experiment.

For Experiment 2, we observed only few significant and relatively low correlations between plausibility ratings across epochs (see Tab. 1), indicating that participants by-and-large did not prefer one rendering method over the other with the speech stimulus used in Experiment 2. Thus, in contrast to Experiment 1, we cannot tell whether participants were even able to discriminate between the two rendering methods. Rather, it appears likely that most participants typically could not detect any clear differences between both rendering methods and, for that reason alone, could not reliably decide which rendering method was more plausible. The individual-subject movement data for each epoch indicate that participants' movements were consistent throughout the experiment (see Figs. S9–S12 in the Supplementary Material [42]), suggesting that the movement became automatic for participants already during training or within the first few trials of the experiment.

The plots in Figure 4 suggest that the results of Experiment 2 have a lower between-subject variance than those of Experiment 1. A Levene's test confirmed that the variances of the results are significantly different [$F(1,30) = 4.70$, $p = .038$]. We consider this and the absence of correlations across epochs discussed above as indication that the female speech test signal used in Experiment 2 elicited fewer perceptual differences between the near- and far-field HRTFs than the noise test signal used in Experiment 1.

4 General discussion

Previous research on near-field HRTFs and the perception of nearby sound sources mainly focused on distance estimation accuracy and the role of near-field cues (mainly the ILD cue) on distance judgments, leading to a variety of partly conflicting results on the contribution of near-field

cues to auditory distance perception (see, e.g., Arend et al. [11] for an overview). However, there is very little research investigating other perceptual aspects of near-field HRTFs, even though, especially with the emerging interest in binaural 6-DoF rendering for real-time VR and AR applications that often have limited resources, it is becoming increasingly important to determine whether simulating physically correct near-field cues is perceptually necessary. To address these questions, we conducted two 2AFC experiments in a 6-DoF multimodal AR experience investigating whether near-field HRTFs provide a more plausible binaural reproduction of nearby sound sources than intensity-scaled far-field HRTFs in dynamic multimodal virtual acoustic scenes.

The results of both experiments show no evidence that near-field cues enhance the plausibility of non-individual anechoic binaural rendering of nearby sound sources in the dynamic multimodal virtual acoustic scene designed for this study. Thus, even though in the present study the chance of perceiving a difference between the near- and far-field conditions was maximized because the multimodal AR experience provided proprioceptive and visual cues that could have conflicted with incorrect auditory cues, performance was on average (Experiment 1) or for almost each individual participant (Experiment 2) close to chance level, yielding average p_{2AFC} values slightly below but not significantly different from chance level. The equality in plausibility of the two compared HRTFs is rather surprising, given that near-field HRTFs lead to a physically more accurate representation of the nearby sound field than intensity-scaled far-field HRTFs and should therefore be perceived as more plausible on the (common) assumption that plausibility is governed by physical accuracy. Overall, the data from both experiments even show a (non-significant) trend toward p_{2AFC} values that are clearly below chance, which means numerous participants perceived the intensity-scaled far-field HRTFs as more plausible than the near-field HRTFs. On the other hand, there were participants in both experiments who favored the near-field condition. In Experiment 1, it was two expert listeners who tended toward the near-field HRTFs. However, two other expert listeners performed near or even below chance. The statistical outlier in Experiment 2, who tended to prefer near-field HRTFs, was not classified as an expert listener. Thus, there is no obvious relationship between listening experience and perceived plausibility of the near-field reproduction in the present study.

In both experiments, preference for near-field renderings did not correlate with the time participants placed the virtual sound source in the lateral region, where it would produce strong ILD cues. Moreover, both experiments did not show any learning or fatigue effects throughout the experiment, as revealed by comparing performance across four epochs.

The analysis in epochs also showed that preferences in terms of plausibility strongly correlated across epochs in Experiment 1, but lower and often non-significant correlations were observed in Experiment 2. Thus, some participants' answers were very consistent throughout the first experiment, i.e., they consistently perceived the near-field HRTFs as more plausible; others consistently preferred the far-field HRTFs. These consistent ratings also imply that these participants must have perceived differences between the two rendering methods. In Experiment 2, most participants performed near chance level in all four epochs, suggesting that they did either not have any preferences or did not even perceive any difference between the two rendering methods – even after extensive exposure to the speech stimulus and availability of clear spatial cues and rich multimodal information. This difference in consistency of ratings between experiments is also reflected in the between-subject variance: compared to Experiment 1, ratings in Experiment 2 exhibit a significantly lower between-subject variance, with individual p_{2AFC} values all closer to chance level.

One reason for this pattern of results could be that the speech signal used in Experiment 2 provided smaller perceptual differences than the noise signal used in Experiment 1, so that participants could not distinguish between the two renderings and therefore each individual participant answered more randomly in Experiment 2. As the low-frequency ILD cues are similarly excited by both test signals, we assume that the different results are because the spectral differences between the near- and far-field HRTFs (low-pass filtering character), which are strongest at higher frequencies, are much more audible for the broadband noise signal than for the speech signal, which has low energy above 8 kHz.

All these findings suggest – much to our surprise – that using near-field HRTFs or simply continuously adapting ILDs as a function of sound source distance, as done in various binaural renderers, does *not* lead (at least in dynamic multimodal environments) to a more plausible rendering of a virtual sound source in anechoic conditions for naive listeners than a simple rendering with intensity-scaled far-field HRTFs.

In a recent study conducted in a 6-DoF VR environment by Rummukainen et al. [14], listeners did not prefer measured multi-distance near-field HRTFs over intensity-scaled far-field HRTFs for non-individual anechoic dynamic binaural near-field rendering. The authors therefore concluded that including near-field HRTFs provides little benefit in a 6-DoF VR environment. However, the closest distance examined in their study was 0.50 m and the distance resolution was low, both because they used a near-field HRTF set measured with a Neumann KU100 at distances of 0.50, 0.75, 1.00, and 1.50 m [5]. As the strongest distance-dependent near-field effects occur below 0.50 m [3, 5], the authors mentioned that further studies with closer distances are necessary to be able to make a conclusion.

With the present study, we made another attempt to investigate whether near-field HRTFs provide an advantage for non-individual binaural reproduction in an anechoic 6-DoF VR or AR environment, but avoided above-mentioned drawbacks by using near-field HRTFs for very close distances down to 0.12 m at a much higher resolution of 1 cm in distance. In general, our results support the (partly inconclusive) findings of Rummukainen

et al. [14] that, from a perceptual point of view, near-field HRTFs provide little to no benefit for naive listeners in 6-DoF VR or AR multimodal applications employing binaural synthesis. In contrast, previous studies such as those by Brungart [6] or Kan et al. [12], claimed that near-field HRTFs are mandatory to generate binaural near-field rendering (based on the general assumption that a physically correct near-field representation is necessary). However, these conclusions are based on studies on distance estimation accuracy, which we did not investigate in our experiments. Thus, the importance of near-field HRTFs might differ depending on the task or application, i.e., if high-precision distance estimation accuracy in the near field is mandatory in an application, near-field HRTFs might provide advantages, whereas our results suggest that they are not necessary for an overall plausible representation of a dynamic spatial sound scene.

Our experiments, as well as the study by Rummukainen et al. [14], might indicate that correct reproduction of intensity as the primary and strongest distance cue is, in most cases, sufficient for a plausible representation of nearby sound sources in dynamic multimodal virtual environments. In line with this, experiments on distance perception revealed that, if available, the intensity cue dominates auditory distance estimation and masks the much more subtle near-field cues [11, 13]. Furthermore, a multimodal AR experience, as in the present study, provides proprioceptive and visual cues in addition to auditory cues, enhancing auditory localization and providing listeners with more information to judge the plausibility of a virtual sound source reliably. Conforming to this, previous studies on auditory space adaptation and multisensory learning effects have found evidence indicating that kinesthetic cues are additive to those evoked when the listener only pays attention to the sound source or can only see its position in space, suggesting that kinesthetic cues further support the spatial hearing updating process [35, 50, 51]. Valzolgher et al. [35], for example, considered that the sensory input achieved by multimodal stimulation, which is also supported by the human intention to act in space, could contribute to tuning the listener's sound-space correspondences. Moreover, similar to the intensity cue, these strong visual and proprioceptive cues might mask the more subtle near-field cues. To summarize, there are two possible effects of multimodal stimulation on plausibility assessment, which may even interact with each other. On the one hand, there is significant scientific evidence that multimodal stimulation combining real and simulated information improves plausibility judgments, as the different information streams can be evaluated concerning their congruency, and possible incoherences between the streams appear immediately as a break in plausibility. On the other hand, simultaneous streams containing real information congruent with the simulated auditory information might mask (in addition to the intensity cues) the less salient near-field cues, probably making the AR experience plausible even with simple distance-dependent intensity-scaling of far-field HRTFs.

The results are of particular relevance for real-time VR and AR applications with limited resources that use (mostly non-individual) binaural synthesis for 6-DoF rendering of virtual sound sources. Our results suggest that the additional (computational) effort of including near-field cues or near-field HRTF synthesis may not be necessary in terms of plausibility and reproduction quality for multimodal scenes. Furthermore, most applications reproduce reverberant environments, in which early reflections and reverberation would most probably further reduce perceptual differences between near- and far-field HRTFs. As our results suggest that even in anechoic environments using near-field HRTFs provides no perceptual benefit in terms of plausibility for naive listeners, we assume that all the more there is no benefit in using near-field HRTFs for reproducing reverberant environments.

Acknowledgments

We are grateful to the three anonymous reviewers for their constructive comments on a previous version of this manuscript. We also give special thanks to all the participants in the study. This work was supported by the German Federal Ministry of Education and Research (03FH014IX5-NarDasS and 13FH666IA6-VIWER-S).

Supplementary material

Supplementary material containing the Matlab script developed to generate the HRTFs and the filtered test signals, a video illustrating the experimental procedure, and additional results figures is available at https://doi.org/10.5281/zenodo.5656726.

Conflict of interest

Authors declared no conflict of interests.

References

1. P. Zahorik, D.S. Brungart, A.W. Bronkhorst: Auditory distance perception in humans: A summary of past and present research. Acta Acustica United with Acustica 91, 3 (2005) 409–420.
2. A.J. Kolarik, B.C.J. Moore, P. Zahorik, S. Cirstea, S. Pardhan: Auditory distance perception in humans: A review of cues, development, neuronal bases, and effects of sensory loss. Attention, Perception, & Psychophysics 78, 2 (2016) 373–395. https://doi.org/10.3758/s13414-015-1015-1.
3. D.S. Brungart, W.M. Rabinowitz: Auditory localization of nearby sources. Head-related transfer functions. The Journal of the Acoustical Society of America 106, 3 (1999) 1465–1479. https://doi.org/10.1121/1.427180.
4. D.S. Brungart, W.M. Rabinowitz: Auditory localization in the near-field, in Proc. of the 3rd International Conference on Auditory Display, Palo Alto, CA, USA. 1996, pp. 1–5.
5. J.M. Arend, A. Neidhardt, C. Pörschmann: Measurement and perceptual evaluation of a spherical near-field HRTF set, in Proc. of the 29th Tonmeistertagung – VDT International Convention, Cologne, Germany. 2016, pp. 356–363.

6. D.S. Brungart: Auditory localization of nearby sources. III. Stimulus effects. The Journal of the Acoustical Society of America 106, 6 (1999) 3589–3602. https://doi.org/10.1121/1.428212.
7. N. Kopčo, B.G. Shinn-Cunningham: Effect of stimulus spectrum on distance perception for nearby sources. The Journal of the Acoustical Society of America 130, 3 (2011) 1530–1541. https://doi.org/10.1121/1.3613705.
8. N. Kopčo, S. Huang, J.W. Belliveau, T. Raij, C. Tengshe, J. Ahveninen: Neuronal representations of distance in human auditory cortex. Proceedings of the National Academy of Sciences 109, 27 (2012) 11019–11024. https://doi.org/10.1073/pnas.1119496109.
9. N. Kopčo, K. Kumar Doreswamy, S. Huang, S. Rossi, J. Ahveninen: Cortical auditory distance representation based on direct-to-reverberant energy ratio. NeuroImage 208 (2020) 116436. https://doi.org/10.1016/j.neuroimage.2019.116436.
10. B.G. Shinn-Cunningham: Localizing sound in rooms, in Proc. of the ACM SIGGRAPH and EUROGRAPHICS Campfire: Acoustic Rendering for Virtual Environments, Snowbird, Utah. 2001, pp. 17–22.
11. J.M. Arend, H.R. Liesefeld, C. Pörschmann: On the influence of non-individual binaural cues and the impact of level normalization on auditory distance estimation of nearby sound sources. Acta Acustica 5, 10 (2021) 1–21. https://doi.org/10.1051/aacus/2021001.
12. A. Kan, C. Jin, A. van Schaik: A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function. The Journal of the Acoustical Society of America 125, 4 (2009) 2233–2242. https://doi.org/10.1121/1.3081395.
13. S. Spagnol, E. Tavazzi, F. Avanzini: Distance rendering and perception of nearby virtual sound sources with a near-field filter model. Applied Acoustics 115 (2017) 61–73. https://doi.org/10.1016/j.apacoust.2016.08.015.
14. O.S. Rummukainen, S.J. Schlecht, T. Robotham, A. Plinge, E.A.P. Habets: Perceptual study of near-field binaural audio rendering in six-degrees-of-freedom virtual reality, in Proc. of IEEE VR, Osaka, Japan. 2019, pp. 1–7. https://doi.org/10.1109/VR.2019.8798177.
15. A. Lindau, S. Weinzierl: Assessing the plausibility of virtual acoustic environments. Acta Acustica United with Acustica 98, 5 (2012) 804–810. https://doi.org/10.3813/AAA.918562.
16. M. Slater: Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philosophical Transactions of the Royal Society B 364 (2009) 3549–3557. https://doi.org/10.1098/rstb.2009.0138.
17. M. Hofer, T. Hartmann, A. Eden, R. Ratan, L. Hahn: The role of plausibility in the experience of spatial presence in virtual environments. Frontiers in Virtual Reality 10, April (2020) 1–9. https://doi.org/10.3389/frvir.2020.00002.
18. U. Reiter: Perceived quality in game audio, in Grimshaw M (Ed.), Game Sound Technology and Player Interaction: Concepts and Developments, Chapter 8, IGI Global, Hershey, PA, USA. 2011, pp. 153–174. https://doi.org/10.4018/978-1-61692-828-5.ch008.
19. D. Ackermann, F. Fiedler, F. Brinkmann, M. Schneider, S. Weinzierl: On the acoustic qualities of dynamic pseudobinaural recordings. The Journal of the Audio Engineering Society 68, 6 (2020) 418–427. https://doi.org/10.17743/jaes.2020.0036.
20. J.M. Arend, S.V. Amengual Garí, C. Schissler, F. Klein, P.W. Robinson: Six-degrees-of-freedom parametric spatial audio based on one monaural room impulse response. The Journal of the Audio Engineering Society 69, 7/8 (2021) 557–575. https://doi.org/10.17743/jaes.2021.0009.
21. F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, S. Weinzierl: A round robin on room acoustical simulation and auralization. The Journal of the Acoustical Society of America 145, 4 (2019) 2746–2760. https://doi.org/10.1121/1.5096178.
22. A. Neidhardt, N. Knoop: Binaural walk-through scenarios with actual self-walking using an HTC Vive, in Proc. of the 43rd DAGA, Kiel, Germany. 2017, pp. 283–286.
23. A. Neidhardt, A.I. Tommy, A.D. Pereppadan: Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets, in Proc. of the 144th AES Convention, Milan, Italy. 2018, pp. 1–11.
24. S.V. Amengual Garí, J.M. Arend, P. Calamia, P.W. Robinson: Optimizations of the spatial decomposition method for binaural reproduction. The Journal of the Audio Engineering Society 68, 12 (2020) 959–976. https://doi.org/10.17743/jaes.2020.0063.
25. A. Neidhardt, A.M. Zerlik: The availability of a hidden real reference affects the plausibility of position-dynamic auditory AR. Frontiers in Virtual Reality 2, 678875 (2021) 1–17. https://doi.org/10.3389/frvir.2021.678875.
26. VRACE: VRACE Research Team. https://vrace-etn.eu/research-team/. Accessed: 2021-11-09.
27. Oculus: Oculus Developer. https://developer.oculus.com/blog/near-field-3d-audio-explained. Accessed: 2021-11-09.
28. Magic Leap: Magic Leap Developer. https://developer.magicleap.com/en-us/learn/guides/lumin-sdk-soundfield-audio. Accessed: 2021-11-09.
29. Resonance Audio: Resonance Audio Developer. https://resonance-audio.github.io/resonance-audio/develop/overview.html. Accessed: 2021-11-09.
30. T. Carpentier, M. Noisternig, O. Warusfel: Twenty years of Ircam Spat: Looking back, looking forward, in Proc. of 41st International Computer Music Conference (ICMC), Denton, TX, USA. 2015, pp. 270–277.
31. D. Poirier-Quinot, B.F.G. Katz: The Anaglyph binaural audio engine, in Proc. of the 144th AES Convention, Milan, Italy. 2018, pp. 1–4.
32. M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, A. Reyes-Lecuona: 3D tune-in toolkit: An open-source library for real-time binaural spatialisation. PLoS One 14, 3 (2019) 1–37. https://doi.org/10.1371/journal.pone.0211899.
33. K. Strelnikov, M. Rosito, P. Barone: Effect of audiovisual training on monaural spatial hearing in horizontal plane. PLoS One 6, 3 (2011) 1–9. https://doi.org/10.1371/journal.pone.0018344.
34. A. Isaiah, T. Vongpaisal, A.J. King, D.E.H. Hartley: Multisensory training improves auditory spatial processing following bilateral cochlear implantation. The Journal of Neuroscience 34, 33 (2014) 11119–11130. https://doi.org/10.1523/JNEUROSCI.4767-13.2014.
35. C. Valzolgher, C. Campus, G. Rabini, M. Gori, F. Pavani: Updating spatial hearing abilities through multisensory and motor cues. Cognition 204 (2020) 104409. https://doi.org/10.1016/j.cognition.2020.104409.
36. A. Neidhardt, F. Klein, N. Knoop, T. Köllmer: Flexible Python tool for dynamic binaural synthesis applications, in Proc. of the 142nd AES Convention, Berlin, Germany. 2017, pp. 1–5.
37. B. Bernschütz: A spherical far field HRIR/HRTF compilation of the Neumann KU 100, in Proc. of the 39th DAGA, Merano, Italy. 2013, pp. 592–595.
38. R.O. Duda, W.L. Martens: Range dependence of the response of a spherical head model. The Journal of the Acoustical Society of America 104, 5 (1998) 3048–3058. https://doi.org/10.1121/1.423886.
39. V. Ralph Algazi, C. Avendano, R.O. Duda: Estimation of a spherical-head model from anthropometry. The Journal of the Audio Engineering Society 49, 6 (2001) 472–479.

40. D. Romblom, B. Cook: Near-Field Compensation for HRTF Processing, in Proc. of the 125th AES Convention, San Francisco, USA. 2008, pp. 1–6.
41. J.M. Arend, C. Pörschmann: Synthesis of near-field HRTFs by directional equalization of far-field datasets, in Proc. of the 45th DAGA, Rostock, Germany. 2019, pp. 1454–1457.
42. J.M. Arend, M. Ramírez, H.R. Liesefeld, C. Pörschmann: Supplementary material for "Do near-field cues enhance the plausibility of non-individual binaural rendering in a dynamic multimodal virtual acoustic scene?". Nov. 2021. https://doi.org/10.5281/zenodo.5656726.
43. A. Lindau, F. Brinkmann: Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings. The Journal of the Audio Engineering Society 60, 1/2 (2012) 54–62.
44. V. Erbes, M. Geier, H. Wierstorf, S. Spors: Free database of low-frequency corrected head-related transfer functions and headphone compensation filters, in Proc. of the 127th AES Convention, New York, NY, USA. 2017, pp. 1–5.
45. S.W. Greenhouse, S. Geisser: On methods in the analysis of profile data. Psychometrika 24, 2 (1959) 95–112. https://doi.org/10.1007/BF02289823.
46. B. Bruya: Effortless attention: A new perspective in the cognitive science of attention and action. MIT Press, Cambridge, MA, 2010. https://doi.org/10.7551/mitpress/9780262013840.001.0001.
47. W. Schneider, R.M. Shiffrin: Controlled and automatic human information processing: I. Detection, search, and attention. Psychological Review 84, 1 (1977) 1–66. https://doi.org/10.1037/0033-295X.84.1.1.
48. P. Demonte: HARVARD speech corpus – audio recording 2019. University of Salford. Collection, 2019. URL https://doi.org/10.17866/rd.salford.c.4437578.v1.
49. ITU-R BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level. International Telecommunications Union, Geneva, 2015.
50. A. Maravita, C. Spence, J. Driver: Multisensory integration and the body schema: Close to hand and within reach. Current Biology 13, 13 (2003) 531–539. https://doi.org/10.1016/S0960-9822(03)00449-4.
51. M. Gori, T. Vercillo, G. Sandini, D. Burr: Tactile feedback improves auditory spatial localization. Frontiers in Psychology 5 (2014) 1–7. https://doi.org/10.3389/fpsyg.2014.01121.

Cite this article as: Arend J. Ramírez M. Liesefeld HR. & Pörschmann C. 2021. Do near-field cues enhance the plausibility of non-individual binaural rendering in a dynamic multimodal virtual acoustic scene?. Acta Acustica, **5**, 55.

4.2 On the Influence of Non-Individual Binaural Cues and the Impact of Level Normalization on Auditory Distance Estimation of Nearby Sound Sources

Arend, J. M., Liesefeld, H. R., & Pörschmann, C. (2021). *Acta Acust.*, *5*(10), 1–21. https://doi.org/10.1051/aacus/2021001

Acta Acustica 2021, **5**, 10

https://doi.org/10.1051/aacus/2021001

Available online at:
https://acta-acustica.edpsciences.org

SCIENTIFIC ARTICLE OPEN ACCESS

On the influence of non-individual binaural cues and the impact of level normalization on auditory distance estimation of nearby sound sources

Johannes M. Arend[1,2,*], Heinrich R. Liesefeld[3,4], and Christoph Pörschmann[1]

[1] Institute of Communications Engineering, TH Köln – University of Applied Sciences, Betzdorfer Str. 2, 50679 Cologne, Germany
[2] Audio Communication Group, Technical University of Berlin, Einsteinufer 17c, 10587 Berlin, Germany
[3] Department of Psychology, University of Bremen, Hochschulring 18, 28359 Bremen, Germany
[4] Department of Psychology, Ludwig-Maximilians-Universität München, Leopoldstr. 13, 80802 Munich, Germany

Received 18 June 2020, Accepted 14 January 2021

Abstract – Nearby sound sources provide distinct binaural cues, mainly in the form of interaural level differences, which vary with respect to distance and azimuth. However, there is a long-standing controversy regarding whether humans can actually utilize binaural cues for distance estimation of nearby sources. Therefore, we conducted three experiments using non-individual binaural synthesis. In Experiment 1, subjects had to estimate the relative distance of loudness-normalized and non-normalized nearby sources in static and dynamic binaural rendering in a multi-stimulus comparison task under anechoic conditions. Loudness normalization was used as a plausible method to compensate for noticeable intensity differences between stimuli. With the employed loudness normalization, nominal distance did not significantly affect distance ratings for most conditions despite the presence of non-individual binaural distance cues. In Experiment 2, subjects had to judge the relative distance between loudness-normalized sources in dynamic binaural rendering in a forced-choice task. Below chance performance in this more sensitive task revealed that the employed loudness normalization strongly affected distance estimation. As this finding indicated a general issue with loudness normalization for studies on relative distance estimation, Experiment 3 directly tested the validity of loudness normalization and a frequently used amplitude normalization. Results showed that both normalization methods lead to remaining (incorrect) intensity cues, which subjects most likely used for relative distance estimation. The experiments revealed that both examined normalization methods have consequential drawbacks. These drawbacks might in parts explain conflicting findings regarding the effectiveness of binaural cues for relative distance estimation in the literature.

1 Introduction

The primary acoustic cues for distance perception in the far field are sound intensity, direct-to-reverberant energy ratio (DRR), and spectral cues [1, 2]. Whereas the DRR cue provides absolute distance information, intensity and spectrum are relative distance cues, which means that different sounds have to be compared in order to judge distance. In reverberant environments, humans can use a combination of these cues for distance estimation, albeit spectral distance cues caused by high frequency attenuation are only available for sound sources with a distance of more than 15 m ([3], Chapter 2.3.2). In anechoic conditions, distance estimation of far field sources (below 15 m) relies mainly on intensity cues.

In the near field (sound source distance less than 1 m), interaural time differences (ITDs), interaural level differences (ILDs), and characteristic spectral cues occur in addition [2, 4]. The spectral properties of nearby sound sources change across distance. Diffraction and head-shadowing effects lead to a low-pass filtering character of nearby sound sources which might be a spectral cue for distance estimation in the near field [1, 2, 4]. While the influence of distance on the ITDs is relatively low [4, 5], ILDs change substantially across distance for nearby sound sources [4, 5].

The increase in ILDs as a sound source approaches the head is mainly caused by frequency-dependent head-shadowing effects and might be the most prominent feature of nearby sound sources. The strongest increase in ILDs can be observed for lateral sound sources at distances below 0.50 m [4]. For example, broadband ILDs obtained from HRTFs (Head-Related Transfer Functions) measured with a dummy head can increase about 10 dB to an order of 20 dB [4] or 23 dB [6] for a lateral sound source at a distance of 0.12 m or 0.25 m respectively. Because of these drastic changes in ILDs across the whole spectrum, it is assumed that ILDs play an important role for distance estimation in the near field [2].

*Corresponding author: Johannes.Arend@th-koeln.de

However, intensity has been shown to be the highest weighted distance cue [7, 8]. To evaluate the contribution of other distance cues, previous research has eliminated intensity differences between stimuli by various kinds of level equalization (see Tab. A.1 in Appendix for an overview). The *normalization method* to equalize levels has a distinct impact on the stimuli. This issue was not considered in detail in previous studies and may explain some conflicting results in the literature as summarized in the following.

Holt and Thurlow [9] conducted an experiment in anechoic conditions with level-equalized sources at distances between 1.80 m and 19 m. Binaural cues were isolated by eliminating DRR (anechoic) and intensity cues (level-equalized). The authors reported that subjects were not able to judge the distance of a frontally oriented sound source presenting a broadband noise stimulus. However, performance improved when the source was positioned laterally. In a similar experiment with level-equalized speech sources arranged in the front at distances between 0.90 m and 9.00 m, Gardner [10] observed that small head movements showed slight benefits for distance estimation. Although both studies addressed far-field sources only, the results indicate that binaural cues might provide additional information for distance estimation. Brungart et al. examined nearby sound sources in two consecutive listening experiments in anechoic conditions [7, 11]. Subjects had to judge the position of a specific sound source (approximating an acoustic point source) randomly located in their right hemifield at distances between 0.15 m and 1.00 m. To remove intensity-based cues and thus to isolate binaural cues, the amplitude of the noise stimulus was normalized dependent on distance. Additionally, to further reduce the reliability of potentially remaining intensity cues, the amplitude of the normalized stimuli was roved randomly over a 15 dB range. The results showed that distance estimation was most accurate for lateral sources and least accurate near the median plane. Moreover, accuracy of distance estimation degraded when frequency components below 3 kHz were absent. Based on these results, the authors concluded that low-frequency ILDs (below 3 kHz) are the primary and most salient cue for distance estimation of nearby lateral sound sources when no intensity cues are available, whereas listeners rely primarily on intensity cues for nearby medial sources where binaural cues are weak. In a follow-up study, Brungart and Simpson [12] conducted a similar experiment by means of a virtual auditory display based on near-field HRTFs of a KEMAR dummy head [4]. Again, subjects had to judge distance to level-equalized and level-roved (10 dB range) virtual noise sources located in their right hemifield at distances between 0.12 m and 1.00 m. In this study, however, a different type of level normalization was applied. Subjects performed worse than in the experiment with a real sound source [7, 11], which according to the authors was most likely due to the use of non-individual HRTFs. However, the authors stated that especially for lateral sound sources, subjects were still able to extract a substantial amount of distance information from the normalized nearby virtual sound sources. Kan et al. [13] also conducted an experiment in virtual acoustics applying near-field HRTFs synthesized from individual far-field HRTFs. Subjects had to judge the distance of virtual noise sources at distances between 0.10 m and 1.00 m. To remove intensity cues, the authors applied the same normalization method as Brungart et al. [7], but without level-roving. The results showed a distance discrimination for lateral sound sources within a range of 0.20 m, but the overall distance judgment performance was poor and the authors concluded that ILDs are no powerful cues. Spagnol et al. [8] conducted a similar study using synthesized and measured KEMAR near-field HRTFs. Subjects were asked to discriminate distance of virtual lateral and medial noise sources at distances between 0.20 m and 1.00 m. The researchers also applied the amplitude normalization proposed by Brungart et al. [7] without level-roving. The experiment resulted in average error rates very close to the 50% chance level, indicating that the overall performance was poor. However, similar to Kan et al. [13], performance slightly improved for lateral sources at distances below 0.20 m.

In contrast to these findings supporting the effectiveness of binaural cues, several other studies cast doubt whether binaural cues contribute to distance estimation. Simpson and Stanton [14] conducted experiments in quasi-anechoic conditions with a pulse-train source located in the front at distances between 0.30 m and 2.66 m and found that head movements had no influence on distance estimation and concluded that binaural cues are unimportant for distance perception. Rosenblum et al. [15] also noted that head movements had no influence on distance judgment accuracy. In their experiments, subjects had to judge the distance to a percussion shaker, located laterally at distances between 0.38 m and 1.10 m. However, since the sound sources were not level-equalized in both of these studies, intensity cues might have masked binaural cues so that no influence of head movements was found. Shinn-Cunningham et al. [16] addressed the question of whether binaural cues contribute to distance estimation in two experiments based on binaural synthesis using individualftable BRIRs (Binaural Room Impulse Responses) and individual HRTFs for reverberant and anechoic conditions respectively. Subjects had to judge distance of medial and lateral virtual pink noise sources at distances between 0.15 m and 1.00 m. Because the performance of untrained listeners was poor both for lateral and medial sound sources in anechoic conditions (distance perception was generally below chance), the authors concluded that binaural cues are weak or even irrelevant. Moreover, they observed that performance in anechoic conditions can improve with training, indicating that listeners can learn to use ILD cues for distance estimation of nearby sound sources. Finally, the authors stated that in reverberant conditions, DRR provides a robust distance cue in the near field, even for medial sound sources and untrained listeners. Based on these experiments, Shinn-Cunningham et al. concluded that ILDs do not contribute to distance estimation when reverberation is present, and that even in anechoic conditions, ILD cues do not lead to robust distance percepts [17, 18]. In a further study, Kopčo and Shinn-Cunningham [19] examined how changes in DRR and ILD affect distance

judgments. Again, the researchers used individual BRIRs to synthesize virtual lateral and medial sound sources at distances between 0.15 m and 1.70 m. To eliminate intensity cues, the noise stimuli were normalized in level, and to further diminish potentially remaining intensity cues, the normalized stimuli were level-roved over a 10 dB range. Similar to previous experiments, the results showed that performance was best for lateral sound sources and worse without low-frequency energy, but the authors concluded that listeners only use DRR cues to judge distance of nearby sound sources in reverberant condition. However, the authors further outlined that listeners might focus on different strategies to judge distance, depending on the listening conditions. In a later study using non-individual BRIRs, Kopčo et al. [20] qualified their statement by saying that listeners might combine the DRR and ILD cue for distance estimation, even though the DRR cue seems to be more robust and reliable than the ILD cue.

Given the conflicting results and despite the drastic variations in ILD induced by distance changes, the contribution of binaural cues to distance estimation in the near field remains an open issue. Even very similar studies strictly focusing on distance perception of nearby sound sources in anechoic conditions, like Brungart and Simpson [12] and Shinn-Cunningham et al. [16] for example, led to opposing conclusions regarding the influence of binaural cues. When comparing all studies, three factors stand out which could have had a significant influence on the respective results: the specific normalization method to eliminate intensity cues, the way head movements were considered, and the use of individual or non-individual HRTFs if binaural synthesis was applied. The following paragraphs shortly discuss those three aspects and their potential influence. Additionally, Table A.1 in Appendix gives an overview of mentioned studies including their method and the major findings concerning the contribution of binaural cues.

1.1 Normalization

One main reason for the differences between the results might be the normalization method. According to the pressure-discrimination hypothesis, just-noticeable differences (JNDs) in source distance are determined by the ability of discriminating changes in source intensity [2, 21]. Considering that the smallest detectable change in sound pressure level is about 0.4 dB for broadband noise [22], it is apparent that specific care has to be taken in the normalization to completely remove intensity cues so that intensity cannot be used instead of binaural cues. Brungart et al. [7], Kan et al. [13], and Spagnol et al. [8] applied the same distance-based normalization (see Sect. 4.1.2 for more details), which only approximately normalizes the amplitude of the stimuli (as acknowledged by Brungart et al. [7]). These three studies have in common that distance discrimination between normalized stimuli was most accurate for lateral sources really close to the head at distances below 0.20 m. In contrast, Kopčo and Shinn-Cunningham [19] normalized the stimuli so that the overall sound pressure level at the nearer ear was constant. The authors could not find any evidence that binaural cues were used for distance estimation in reverberant conditions. In previous studies, Shinn-Cunningham and Kopčo found similar results also for anechoic conditions, but the authors did not provide detailed information on their normalization method [16–18]. The comparison shows that the normalization method may significantly affect the results, for example whether intensity cues remain even after normalization. Additionally roving the stimuli in level after normalization, as for example done by Brungart et al. [7, 11] and Kopčo and Shinn-Cunningham [19], further diminishes potentially remaining intensity cues. However, roving seems not expedient for experiments on relative distance estimation, that is, experiments where at least two sound sources are presented concurrently or in quick succession, and relative distance differences have to be rated. In this case, roving would negate the normalization and introduce intensity cues that most certainly would dominate relative distance estimation (i.e., responses would likely almost exclusively be based on the level differences (re-)introduced by the roving procedure). This problem of level roving was demonstrated in a recent study from Prud'homme and Lavandier [23], where naive listeners judged distance primarily based on the roving-induced level variations, even though they were instructed to discard them. It is therefore necessary to find a reliable normalization method for studies on relative distance estimation. The performance of the various normalization methods or their possible impact on the results in studies of the type discussed above has not been systematically examined so far.

Instead of simply matching the levels of the stimuli in some way, normalizing the stimuli in loudness according to ITU-RBS.1770-4 [24] might be a better approach to remove intensity cues. In comparison to a strictly level-based analysis, loudness is a psychoacoustic measure that takes into account the frequency-dependent sensitivity of the human ear as well as the acoustic effects of the human head. In the course of development, the ITU evaluated the performance of the loudness algorithm in several listening experiments. These experiments yielded correlation coefficients of about $r = 0.98$ between perceived (subjective) loudness measurements and (objective) predicted loudness for a broad range of signals. Thus, the loudness model performs well and normalizing the stimuli in loudness should, arguably, work considerably better than previously introduced purely technical normalization methods, which do not consider the effects of human perception in the same way.

1.2 Head movements

Head movements have been shown to be another important factor influencing distance perception of nearby sound sources. Gardner [10] for example found that the small changes in binaural cues caused by head movements slightly improved distance estimation. In contrast, Simpson and Stanton [14] and Rosenblum et al. [15] reported no influence of head movements. Nevertheless, all studies outlined in this article neither captured head movements for further post-hoc analysis nor considered head movements

if binaural synthesis was applied. However, especially for nearby sound sources, head movements in the horizontal plane lead to distinct variations in ILDs. These variations might provide additional information for distance estimation in the near field and thus probably enhance human perception of auditory space, quite similar to the observation that localization performance in the horizontal and median plane improves if head movements are involved [25]. Therefore, it seems important to analyze the extent of head movements and how they influence distance estimation.

1.3 HRTFs

In comparison to non-individual HRTFs, individual HRTFs improve localization in the median plane and lead to reduced front-back confusion in static binaural synthesis because of more accurate monaural spectral cues [26–28]. In contrast, non-individual HRTFs still provide robust binaural cues for localization in the horizontal plane [26], often without relevant increase in localization error when compared to individual HRTFs [28]. However, it is not clear whether individual HRTFs lead to more accurate distance perception than non-individual HRTFs. Zahorik showed that in reverberant conditions, the performance of distance estimation of virtual sound sources is unaffected by the use of non-individualized HRTFs compared to the use of individualized HRTFs, since the strong intensity and DRR cues mask spectral deviations [1, 29, 30]. In line, Begault et al. [28] did not find any effects of individual HRTFs in anechoic and reverberant conditions on externalization, which is a perceptual attribute associated with distance perception [31]. Likewise, Yu et al. found no evidence for an influence of individual HRTFs on distance perception of nearby sound sources [32, 33]. Most recently, Prud'homme and Lavandier [23] showed that the use of non-individual BRIRs instead of individual BRIRs did not significantly affect absolute distance estimation in reverberant conditions, which confirmed Zahorik's findings. However, Hartmann and Wittenberg [34], Brimijoin et al. [35], or Baumgartner et al. [36] showed that spectral cues affect externalization, and that distorted spectral cues, such as those from non-individual HRTFs, reduce externalization and therefore perceived distance.

Thus, whereas individual HRTFs improve performance especially in median plane localization, their role in distance estimation is not entirely clear [31]. This is also evident when looking at previous studies that provide contradictory results that cannot be directly attributed to the type of HRTFs used. For example, Shinn-Cunningham et al. [16], Shinn-Cunningham [18], and Kopčo and Shinn-Cunningham [19] used individual HRTFs and could not find any evidence that binaural cues contribute to auditory distance estimation of nearby sound sources. In contrast, Brungart and Simpson [12] used generic KEMAR HRTFs and confirmed the findings from their earlier loudspeaker-based study. Brungart and Simpson [12] discussed the findings from Shinn-Cunningham et al. [16] in their article, but could not find a conclusive explanation for the conflicting results. Kopčo et al. [20] used non-individual HRTFs and found that the DRR cue masks potential ILD cues. Again in contrast, Kan et al. [13] and Spagnol et al. [8] both found a slight improvement in distance estimation performance for lateral close sources using individual or non-individual HRTFs respectively. Taken together, regardless of whether individual or non-individual HRTFs were used, the various studies led to contrary results and no direct correlation between the type of HRTFs and the respective findings can be found (see Tab. A.1 in Appendix). Rather, it seems that other factors, such as the test paradigm or the normalization method, have a greater influence than the HRTFs.

1.4 The current study

The detailed review of the studies in this field reveals a long-standing controversy and shows that the question of whether binaural cues contribute to distance perception is still an open issue. To address this, we conducted three listening experiments investigating distance perception of nearby virtual sound sources in anechoic conditions using non-individual binaural synthesis. Experiment 1 (Sect. 2) based on a multi-stimulus comparison method where subjects had to rate perceived distance to loudness-normalized (according to ITU-RBS.1770-4 [24]) and non-normalized stimuli in static or dynamic binaural rendering. In Experiment 2 (Sect. 3), we conducted a relative perceptual distance experiment between loudness-normalized nearby virtual sound sources. As Experiment 2 provided some ambiguous results, we conducted Experiment 3 (Sect. 4) as a follow-up to examine the performance of the loudness normalization and of the amplitude normalization proposed by Brungart et al. [7]. In particular, subjects of Experiment 3 rated the relative perceived loudness-difference between normalized nearby virtual sound sources.

2 Experiment 1

By asking listeners to estimate auditory distance to normalized nearby virtual sound sources in static or dynamic non-individual binaural synthesis, we tested whether distance estimation of nearby sound sources is possible without intensity cues. The particular goals were to test whether distance-related changes in binaural cues may be utilized to distinguish distance and whether head movements and the resulting variations in ILD provide additional information improving distance estimation. In another experimental condition, we maintained the distance-related level differences. The aim of this experimental condition was to examine whether head movements improve distance estimation even when intensity cues are provided, or whether intensity simply masks any additional binaural information.

2.1 Method

2.1.1 Participants

In total, 50 adults took part in the experiment for monetary remuneration (10 Euro per hour). Most of them were students in media technology or electrical engineering.

The participants were divided into two equal groups, with one group performing with head tracking (hereafter abbreviated as group *head tracking* – HT) and the other group performing without head tracking, i.e., using static binaural synthesis (hereafter abbreviated as group *static* – ST). Group HT was composed of 20 males and 5 females aged between 18 and 30 years ($M = 24.44$ years, $Mdn = 25$ years, $SD = 3.06$). Twelve participants of this group (48%) had already taken part in previous listening experiments and thus were familiar with the binaural reproduction system. Group ST was composed of 18 males and 7 females with an age between 19 and 31 years ($M = 23.76$ years, $Mdn = 22$ years, $SD = 3.28$). Here, 7 participants (28%) already had gained experience in former listening tests. However, all participants were naive as to the purpose of this experiment. Moreover, there was no previous training in distance estimation of nearby sound sources, which means that the participants had to rely on their life experience in perceiving nearby sound sources. All participants reported normal hearing.

2.1.2 Setup and stimuli

2.1.2.1 Setup

The experiment took place in the anechoic chamber of TH Köln, which has a low background noise of about 20 dB (A). The participants sat on an office swivel chair so that they could turn easily. The entire experiment was implemented, controlled, and executed with the MATLAB based software Scale [37], running on an Apple iMac. Scale handled the playback of the anechoic audio test signals as well as the internal audio routing in combination with the JACK Audio Connection Kit. For (dynamic) binaural rendering, the SoundScape Renderer [38] paired with a Fastrak head tracking system at a 120 Hz sampling rate was used. Only rotational head movements in the horizontal plane were considered, whereas vertical or translational head movements were disregarded. Via internal TCP/IP sockets, Scale controlled the renderer to switch between datasets or to change the settings according to the respective test condition. Once a second, Scale saved the head tracking data for further analysis of the head movements. The participants gave their response using an Apple iPad 2 tablet, which mirrored the graphical user interface (GUI) of Scale. The binaural audio signal was converted and amplified with an Fireface UFX audio interface and presented over AKG K601 headphones. The interface was set to a buffer size of 512 samples and a sampling rate of 48 kHz.

2.1.2.2 Test signal

As anechoic test signal, we used a pink noise burst sequence with a burst length of 1.50 s (including 10 ms cosine-squared onset/offset ramps) and an interstimulus interval of 0.50 s. Generally, a broadband signal ensures best possible localization performance ([3], Chapters 2.1 and 2.3). Concerning the special case of nearby sound sources, accurate distance judgment requires low frequency components below 3 kHz [11] or at least at around 300 Hz [19]. Hence, a pink noise signal is a good choice for this

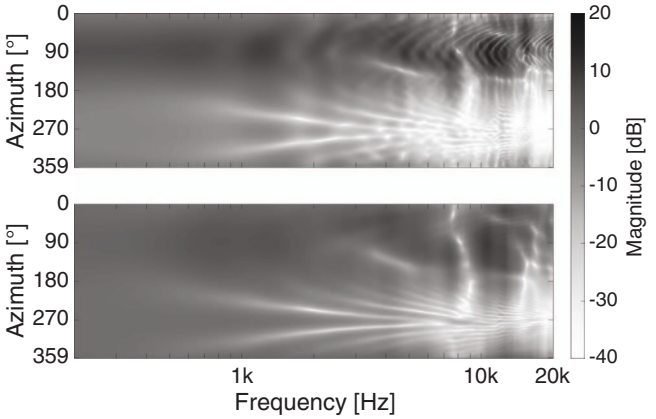


Figure 1. Left-ear magnitude spectrum of two circular grid HRTF datasets (top: $d = 0.25$ m, bottom: $d = 1.50$ m) as a function of frequency (abscissa) and source azimuth φ (ordinate). Comparing the spectrum of near- and far-field HRTFs reveals how a nearby source (top) leads to stronger high-frequency damping around the contralateral ear ($\varphi = 270°$) and increased magnitude around the ipsilateral ear ($\varphi = 90°$), also at frequencies below 3 kHz. This results in significantly higher ILDs (see Fig. 1, left).

experiment. We chose a rather long test signal to provide sufficient time for turning the head during stimulus presentation (as determined during pilot tests) to allow for head-turning related variation in binaural cues. The number of bursts played varied across experiments and is therefore described in the respective procedure paragraph.

2.1.2.3 Near-field HRTFs

To synthesize the nearby virtual sound sources, we used near-field HRTFs from a Neumann KU100 dummy head, measured at five sound source distances ($d = 0.25$ m, 0.50 m, 0.75 m, 1.00 m, 1.50 m) on a circular grid with a resolution of 1° in the horizontal plane [6, 39]. Figure 1 exemplarily shows the left-ear magnitude spectrum of two HRTF datasets (top: $d = 0.25$ m, bottom: $d = 1.50$ m) between 200 Hz and 20 kHz as a function of source azimuth (hereinafter termed direction, or simply φ). Comparing these two extremes (near field vs. far field) clearly reveals how a nearby sound source leads to stronger damping around the contralateral ear (with respect to the sound source, $\varphi = 270°$) and increased magnitude around the ipsilateral ear ($\varphi = 90°$). Moreover, decreased or increased magnitude towards the contralateral and ipsilateral ear respectively can be observed also at frequencies below 3 kHz. Consequently, the ILDs for nearby sources are distinctly higher compared to sources in the far field, as can be seen in Figure 2 (left), which shows the low-frequency ILDs ($f \leq 3$ kHz) of the HRTFs for the five distances. The polar plot shows that low-frequency ILDs, which are considered the primary cue for distance estimation of nearby lateral sources (see Sect. 1), are similar at the distances 1.50 m, 1.00 m, and 0.75 m, start to increase at a distance of 0.50 m, and rise strongly at the closest distance of 0.25 m. Same as for the ILDs, the low-frequency ITDs

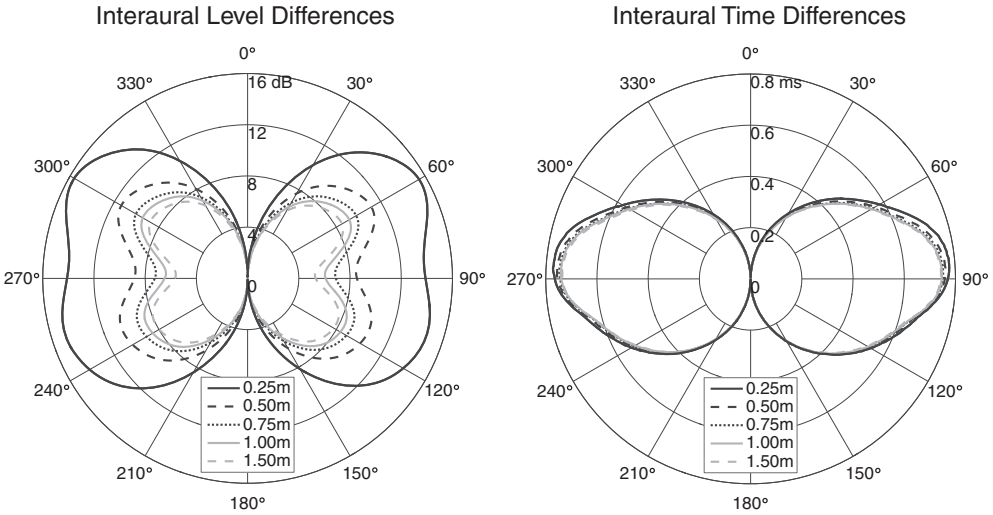


Figure 2. Low-frequency ($f \leq 3$ kHz) interaural level differences (left) and interaural time differences (right) of the used HRTF dataset. The angle represents the azimuth of the sound source (φ). The radius describes the magnitude of the level differences (in dB) or time differences (in ms). Both plots show the usual direction-dependent influence of the pinna and the head. However, typical for near-field HRTFs, the ILDs (left) increase significantly as a function of distance, whereas the ITDs (right) remain nearly constant.

presented in Figure 2 (right) show the usual direction-dependent influence of the pinna and the head. However, with only a slight increase as the source approaches the head, the ITDs are barely influenced by sound source distance. Overall, the analysis presented in this paper as well as the more detailed technical evaluation in Arend et al. [6] and Pörschmann et al. [39] confirm that the HRTFs have the intended near-field characteristics (see, e.g., [2]).

2.1.2.4 Stimuli

In the experiment, virtual noise sources in the horizontal plane (elevation $\vartheta = 0°$) at five different distances (0.25 m, 0.50 m, 0.75 m, 1.00 m, 1.50 m) and three different azimuthal positions (30°, 150°, 270°) per distance were presented, resulting in 15 nominal sound source positions. The positions were chosen to use stimuli covering the front and back hemisphere and showing rather low as well as very distinct binaural cues.

For the non-normalized stimuli, strictly distance-dependent gain values (i.e., independent of the azimuthal position) were calculated in order to maintain natural distance-related level differences. For the normalized stimuli, the loudness of each stimulus was determined according to ITU-RBS.1770-4 for frontal head orientation. For each azimuthal position, the actual determined stimulus loudness at the distance of 1.00 m was set as reference. As a result, stimuli loudness of the normalized stimuli was the same for different distances for one specific azimuthal position, but still varied with respect to the different source azimuths. This choice of a different reference loudness per source azimuth resulted in a slightly different stimuli loudness dependent on azimuth. Thus, for $\varphi = 150°$, overall stimuli loudness across distances was about 1.44 dB LKFS (Loudness, K-weighted, relative to full scale) lower than for $\varphi = 30°$, and for $\varphi = 270°$, overall stimuli loudness across distances was about 1.15 dB LKFS higher than for $\varphi = 30°$. Overall, the procedure resulted in 15 gain values for the loudness-normalized stimuli, as these values were dependent on both distance and azimuth, and 5 gain values for the non-normalized stimuli, as these values were only dependent on distance. Each gain value was assigned to the corresponding virtual sound source in the scene description file of the SoundScape Renderer, thus the actual leveling was applied to the convolution result by the renderer. A double-check with a digital audio workstation metering plugin determining loudness according to EBU R128 [40] confirmed equal loudness for all normalized stimuli.

In order to equalize the binaural chain, a headphone compensation filter according to Bernschütz ([41], Chapter 4.3.4) was applied to the pink noise test signal. The filter was a minimum phase FIR filter with 2048 filter taps. By applying this compensation filter, both the magnitude response of the Neumann KU100 – AKG K601 chain was equalized, and also the loudness normalization was maintained, because a non-flat magnitude response of the reproduction system would have affected perceived loudness at the listener's ear. The (equalized) test signal and the HRTFs (or in this case more precisely the corresponding Head Related Impulse Responses [HRIRs]) were all stored on the control computer as 16 bit/48 kHz .wav files. The playback level for the loudness-normalized stimuli was at about $L_{\mathrm{Aeq}} = 61$ dB. For the non-normalized stimuli, this playback level was assigned to a distance of 1.00 m, resulting in a maximum playback level of about $L_{\mathrm{Aeq}} = 79$ dB for the closest distance of 0.25 m ($\varphi = 270°$).

2.1.3 Procedure

We conducted the experiment with naive listeners only. Pilot tests with the normalized stimuli showed strong learning effects: First, test persons could not immediately distinguish between distances, but when they were given

detailed feedback, they learned to differentiate based on spectral changes, varying ILDs, and head movements. However, since our aim was to examine which cues influence natural distance perception in the near field, we only gave basic instructions about the procedure and refrained from a training session or a scale anchoring process.

The experiment was a 2 × 5 × 3 × 2 mixed factorial design with the between-subjects factor *head tracking* (head tracking, static) and the within-subjects factors *distance* (0.25 m, 0.50 m, 0.75 m, 1.00 m, 1.50 m), source *azimuth* (30°, 150°, 270°), and *normalization* (loudness normalization, no loudness normalization and thus distance-related level differences). We decided to use a mixed design instead of a pure within-subjects design because we observed in pilot tests that participants barely moved their head when head tracking was used as a within-subjects factor, even if we encouraged them to do so. It seemed that the random switch between dynamic and static conditions was quite confusing and distracted them from their actual task, which is maybe why they kept their head still. Further pilot tests with head tracking as a between-subjects factor worked as expected.

Each participant had to attend two separate sessions. In the first session, participants had to rate the normalized, in the second session the non-normalized stimuli. In each session, every participant had to rate the five distances for the three different source azimuths, leading to the 5 × 3 × 2 within-subjects factorial design per group.

The perceived distance had to be rated on a continuous scale with 7 anchor points ("very close", "close", "rather close", "medium", "rather distant", "distant", "very distant") in form of a multi-stimulus comparison method. The same and similar scales for ratings of relative perceived distance have already been successfully used in earlier experiments on distance perception [14, 42, 43]. The procedure was as follows. For each trial, a GUI with five value faders ranging from "very close" to "very distant" was displayed on the tablet. Each fader corresponded to one of the five actual measured distances. The source azimuth was the same for all distances (or faders) within a trial. By touching the respective fader, the participants were able to switch between the corresponding stimuli as often as required, thus also allowing for a comparison of the various stimuli (distances). Technically speaking, the HRTF filter-set switched when touching the fader while the pink noise burst sequence was played in a loop. The order of the faders per trial as well as the order of the trials itself were randomized. The procedure was repeated 10 times per azimuth, thus a full run consisted of 30 trials (with five distance ratings per trial).

Participants of both groups were given the exact same instructions. Regardless of whether they performed the experiment with or without head tracking, they were encouraged to move their head during the estimation process in the form of common localization movements, especially if they felt that movements would improve distance perception. However, they had to keep their front viewing direction because of the different source azimuths. In total, each session lasted for about one hour, including the verbal instruction and a short break.

2.1.4 Data analysis

The statistical analysis was based on the mean values per subject, thus the 10 repetitions per subject for each condition were averaged first. A Jarque-Bera test for normality failed to reject the null hypothesis for 45 out of 60 conditions at a significance level of 0.05. With Hochberg correction [44], which is a common method to correct for multiple hypothesis testing, the test failed to reject the null for all conditions. As parametric tests like the ANOVA are generally robust to slight violations of normality assumptions [45], we analyzed the data using a Greenhouse-Geisser (GG) corrected [46] four-way mixed ANOVA with the between-subjects factor head tracking and the within-subjects factors distance, azimuth, and normalization. For a more detailed analysis, several nested (GG-corrected) mixed and repeated measures ANOVAs as well as paired and independent-samples t tests (two-tailed) at a 0.05 significance level were performed on subsets of the data.

As this analysis revealed no significant effects of head tracking and nominal distance (see the results below), we further analyzed the data using Bayes factors (BF, here BF_{01}). In contrast to common null-hypothesis significance testing, BFs allow stating evidence in favor of the null hypothesis [47, 48]. In particular, the reported BFs are based on independent-samples t tests or nested repeated measures ANOVAs for all effects of additional importance. In brief, BF_{01} expresses the likelihood of the null hypothesis relative to the likelihood of the alternative hypothesis given the data. Thus, for a example, a $BF_{01} = 3$ would suggest that the data provide three times as much evidence for the null than for the alternative. The Bayesian t tests were conducted according to Rouder et al. [48], using the Jeffrey-Zellner-Siow (JZS) prior with a scaling factor of $r = .707$. The BF for the repeated measures ANOVA was calculated according to Rouder et al. [49] using the same prior assumptions.

2.2 Results

Figure 3 shows the results of Experiment 1. The data are separated with respect to the between-subjects factor head tracking and the within-subjects factor normalization, resulting in four subsets: Head Tracking – Loudness Normalization (HTNorm), Head Tracking (HT), Static – Loudness Normalization (STNorm), and Static (ST).

The mean plots in Figure 3 (left) show three notable patterns, statistically confirmed by the analysis further below: (a) There is no apparent effect of head tracking on estimated distance, which can be seen by comparing dynamic (HTNorm, HT) and static (STNorm, ST) conditions. This indicates that head movements had no significant influence on distance estimation of nearby virtual sound sources. (b) Participants did not accurately rate distance of the normalized stimuli (HTNorm and STNorm), suggesting that they could not exploit the non-individual binaural distance cues with the applied loudness normalization method. (c) As expected, participants rated according to the nominal (i.e., actually measured) distance if the

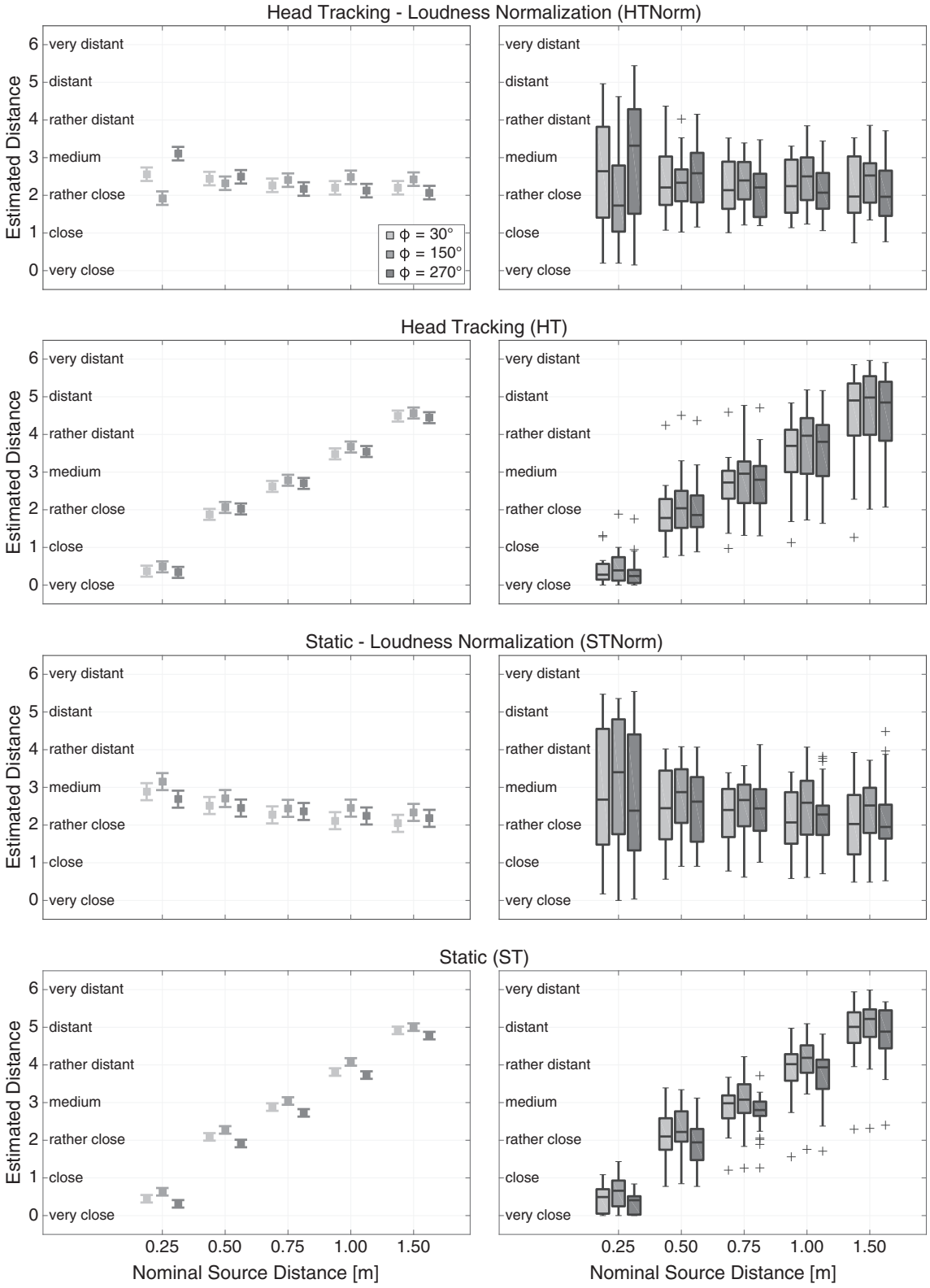


Figure 3. Mean estimated distances (left) and interindividual variation in the estimated distances (right) as a function of nominal source distance (abscissa) and nominal source azimuth (shades of gray) for the subsets: Head Tracking – Loudness Normalization (HTNorm), Head Tracking (HT), Static – Loudness Normalization (STNorm), and Static (ST). The error bars in the mean plots (left) display 95% within-subjects confidence intervals [50, 51], based on the error term of the respective distance main effect. The box plots (right) show the median and the (across participants) interquartile range (IQR) per condition; whiskers display 1.5 × IQR below the 25th or above the 75th percentile and outliers are indicated by plus signs.

Table 1. Results of the four-way mixed ANOVA with the between-subjects factor head tracking (HT) and the within-subjects factors distance (Dist), azimuth (Az), and normalization (Norm).

Source	*df*	*F*	*MSE*	ε	η_p^2	*p*
Between-subjects						
HT	1, 48	.88	11.06	–	.02	.35
Within-subjects						
Dist	4, 192	149.16	1.03	.32	.76	<.001*
Dist × HT	4, 192	.12	1.03	.32	0	.794
Az	2, 96	9.82	.33	.98	.17	<.001*
Az × HT	2, 96	6.56	.33	.98	.12	.002*
Norm	1, 48	15.54	2.71	1	.25	<.001*
Norm × HT	1, 48	.36	2.71	1	.01	.551
Dist × Az	8, 384	5.59	.08	.43	.10	.001*
Dist × Az × HT	8, 384	12.13	.08	.43	.20	<.001*
Dist × Norm	4, 192	219.68	1.15	.31	.82	<.001*
Dist × Norm × HT	4, 192	1.43	1.15	.31	.03	.241
Az × Norm	2, 96	2.37	.24	.97	.05	.101
Az × Norm × HT	2, 96	1.92	.24	.97	.04	.153
Dist × Az × Norm	8, 384	7.55	.07	.38	.14	<.001*
Dist × Az × Norm × HT	8, 384	14.31	.07	.38	.23	<.001*

Note. ε = Greenhouse-Geisser (GG) epsilon, *p* = GG-corrected *p*-values. Note that GG correction is appropriate only for within-subject tests, with more than one degree of freedom in the numerator.
*$p < .05$.

stimuli were not normalized in loudness (HT and ST) and thus natural distance-related intensity cues were provided. Also worth noting are the ratings for HTNorm at $d = 0.25$ m, since participants rated especially the stimuli with $\varphi = 270°$ (erroneously) further away than all other sources. Moreover, the box plots in Figure 3 (right) reveal that the between-subjects variance of the normalized stimuli (especially at $d = 0.25$ m) is considerably higher than for the non-normalized stimuli. This suggests that the normalized conditions might have provided conflicting and ambiguous distance cues that were interpreted or weighted differently by different individuals. In particular, participants might have been confused by binaural cues indicating a nearby sound source in the absence of matching intensity cues.

Table 1 shows the results of the GG-corrected four-way mixed ANOVA. In line with observation (a) made on the basis of the plots, no significant between-subjects effect of head tracking was found, which is also reflected by an independent-samples *t* test on data averaged across all within-subject conditions [$t_{\text{Groups}}(48) = 0.94$, $p = .352$, $d = .27$, $BF_{01} = 2.46$]. This was further confirmed by two independent-samples *t* tests separately testing subsets with and without normalization for an effect of head tracking (HTNorm vs. STNorm, HT vs. ST, averaged across all remaining factors). Both tests yielded no significant difference between the subsets [$t(48) = 0.53$, $p = .60$, $d = .15$, $BF_{01} = 3.15$]; [$t(48) = 1.23$, $p = .227$, $d = .35$ $BF_{01} = 1.92$]. Based on the BFs, the data provided about 2–3 times more evidence for the null than for the alternative, indicating that head tracking did not influence distance estimation performance whether loudness was normalized or not.

The mixed ANOVA revealed several significant within-subjects main and interaction effects though, like for example a rather complex significant four-way interaction between distance, azimuth, normalization, and head tracking. For a better interpretation of the results, we therefore analyzed the data using several nested ANOVAs. In particular, we conducted a GG-corrected two-way repeated measures ANOVA with the within-subjects factors distance and azimuth for each subset. Table 2 summarizes the results of these ANOVAs.

In line with observation (b) made on the basis of the plots, the standard repeated measures ANOVA showed no significant distance effect in the subset HTNorm. For this effect, the respective Bayesian ANOVA revealed $BF_{01} = 5.10$, suggesting that the data of subset HTNorm provide about five times more evidence for the absence (rather than presence) of an effect of distance. For subset STNorm, the main effect of distance from the standard ANOVA was not significant. The Bayesian ANOVA however provided 29,783 more evidence for the presence (rather than absence) of an effect of distance ($BF_{01} = 3.36 \times 10^{-5}$). As evident in Figure 3, this main effect reflected a negative trend of distance, that is, sources that were nominally further away were perceived as closer.

Furthermore, the results of the repeated measures ANOVAs yielded a strong distance × azimuth interaction in subset HTNorm, with an effect of azimuth present mainly for $d = 0.25$ m and to a minor degree for $d = 0.50$ m, and largely absent for the other distances, as revealed by means of further nested ANOVAs (for the sake of conciseness, these one-way repeated measures ANOVAs are not reported here). A paired *t* test yielded a significant difference between the ratings for $d = 0.25$ m and $d = 0.50$ m with $\varphi = 270°$ [$t(24) = 2.67$, $p = .013$, $d_z = .53$], confirming that participants rated the closest source with $\varphi = 270°$ further away than all the other sources. A comparison between the ratings for $d = 0.25$ m and $d = 0.50$ m with $\varphi = 150°$ or $\varphi = 30°$ using paired *t* tests yielded no significant

Table 2. Results of the two-way repeated measures ANOVAs for the subsets HTNorm, HT, STNorm, ST, each with the within-subjects factors distance (Dist) and azimuth (Az).

Source	*df*	*F*	*MSE*	ε	η_p^2	*p*
HTNorm						
Dist	4, 96	.96	1.22	.29	.04	.347
Az	2, 48	.45	.49	.96	.02	.633
Dist × Az	8, 192	18.21	.15	.27	.43	<.001*
HT						
Dist	4, 96	226.87	.81	.34	.90	<.001*
Az	2, 48	6.86	.10	.86	.22	<.001*
Dist × Az	8, 192	1.75	.02	.63	.07	.126
STNorm						
Dist	4, 96	3.21	1.94	.27	.12	.083
Az	2, 48	5.79	.42	.89	.19	.008*
Dist × Az	8, 192	1.87	.10	.38	.07	.142
ST						
Dist	4, 96	570.60	.38	.37	.96	<.001*
Az	2, 48	24.15	.13	.99	.50	<.001*
Dist × Az	8, 192	1.34	.03	.54	.05	.259

Note. ε = Greenhouse-Geisser (GG) epsilon, *p* = GG-corrected *p*-values. Note that GG correction is appropriate only for within-subject tests, with more than one degree of freedom in the numerator.
*$p < .05$.

differences [$t(24) = 2.02$, $p = .055$, $d_z = .40$], [$t(24) = 0.49$, $p = .626$, $d_z = .10$], that is, in subset HTNorm, the source at $d = 0.25$ m and $\varphi = 270°$ was the only source rated significantly different in distance compared to all other sources.

Moreover, there was a main effect of azimuth in subset STNorm without a significant distance × azimuth interaction, but further nested ANOVAs for this subset showed significant azimuth effects only for conditions with $d = 0.25$ m and $d = 1.00$ m, indicating that the influence of source azimuth on estimated distance is relatively small for the remaining levels of distance in this subset.

For subset HT and ST, the ANOVAs revealed highly significant main effects of distance, confirming observation (c) made on the basis of the plots. As expected, participants were able to distinguish distance for the non-normalized stimuli, thus validating the employed procedure. Moreover, the results yielded a rather strong main effect of azimuth for both subsets. These effects can also be seen in the plots, as the means vary in some kind of triangular pattern with respect to azimuth.

To analyze the three-way interaction between distance, azimuth, and head tracking, as well as the two-way interaction between azimuth and head tracking, we first compared the conditions with normalized stimuli (subsets HTNorm and STNorm) as a function of head tracking. Two-way mixed ANOVAs for each level of distance with the between-subjects factor head tracking and the within-subjects factor azimuth showed a highly significant azimuth × head tracking interaction for conditions with $d = 0.25$ m, and a small significant interaction for conditions with $d = 0.50$ m. The plots in Figure 3 clearly illustrate this interaction effect, as the values at $d = 0.25$ m and $d = 0.50$ m vary significantly dependent on head tracking and show opposing patterns (compare HTNorm and STNorm). Similar ANOVAs for the subsets HT and ST revealed a significant azimuth × head tracking interaction for all levels of distance except for $d = 1.50$ m. A look at the plots clarifies this interaction effect, since especially the means at $\varphi = 270°$ are notably higher for conditions with head tracking than for conditions without head tracking.

To further unpack the four-way interaction between distance, azimuth, normalization, and head tracking, we conducted two-way nested mixed ANOVAs with the between-subjects factor head tracking and the within-subjects factor normalization for each combination of levels of the factors distance and azimuth. Here, the results showed a slightly significant interaction between normalization and head tracking for only two conditions ($d = 1.50$ m, $\varphi = 30°$ and $d = 0.25$ m, $\varphi = 150°$), suggesting a rather small influence of the normalization × head tracking interaction in the context of the entire dataset.

To validate that the option to move the head was actually used more often when it had an effect on the stimulation, we also compared the head movements between groups with and without head tracking (groups HT and ST). For each participant, the standard deviation of the horizontal viewing directions (azimuth) for all conditions was calculated, leading to 25 values per group representing the amount of variation around the viewing direction of 0° (front viewing direction). The averaged standard deviation among the group with head tracking was 23.98° and 13.56° among the group without head tracking. An independent-samples *t* test revealed a significant difference between the two groups [$t(48) = 3.39$, $p = .001$, $d = .96$], indicating that the group with head tracking moved their head to a significantly higher degree.

2.3 Discussion

The most interesting results of Experiment 1 are that nominal distance did not significantly affect distance ratings for the normalized stimuli (except for the condition $d = 0.25$ m, $\varphi = 270°$ in the subset HTNorm) and that adapting stimuli to the current head position (head tracking) had no significant influence on estimated distance even though it influenced the degree to which participants moved their heads. The findings indicate that in most conditions, the naive listeners did not use the variations in binaural cues, whether induced by a change in nominal distance of the virtual sound source, or by a change in head orientation. The non-significant distance effect as well as the estimated Bayes factor $BF_{01} = 5.10$ for this effect for conditions with normalized stimuli and dynamic binaural rendering (subset HTNorm) generally support this assumption. Only the significant distance × azimuth interaction in the same subset, mainly driven by an effect of azimuth for $d = 0.25$ m, could be attributed to an effect of binaural cues. Surprisingly, however, the most nearby source with $\varphi = 270°$ for this distance in subset HTNorm was rated as being the furthest away. Furthermore, for subset STNorm, perceived distance decreased with nominal distance. These two, at first sight counterintuitive, findings can be better explained by the workings of intensity cues rather than binaural cues, as revealed by Experiment 3 (see Sect. 4).

The findings regarding the importance of binaural cues conflict with the well know results from Brungart et al. [7, 11, 12], who concluded that especially low-frequency ILDs are an important binaural distance cue for nearby sound sources in real or virtual acoustics. Even though the experiments differ in many ways, like for example the general setup and procedure (e.g., Brungart et al. [17] roved the level of the stimuli in their experiment on absolute distance judgments, in order to diminish potentially remaining intensity cues), the employed HRTFs, or the applied binaural synthesis (Brungart and Simpson [12] employed KEMAR HRTFs and only used static binaural synthesis), it is not directly obvious why the findings are so different. It appears possible that the test procedure and the method to eliminate intensity cues in the stimuli have a major impact on the results (see Sect. 5 for a detailed discussion).

The observed main effect of azimuth can, to some extent, be explained by the slight differences in stimuli loudness dependent on the azimuthal position, as described in Section 2.1.2. The data show a triangular pattern as a function of azimuth (see Fig. 3), except for the above-discussed conditions $d = 0.25$ m and $d = 0.50$ m in subset HTNorm, where this pattern appears inverse, leading to the distance × azimuth interaction effect in subset HTNorm. Thus, in most cases, participants rated conditions with $\varphi = 150°$ a little bit further away than the stimuli with $\varphi = 30°$, most likely because the stimuli with $\varphi = 150°$ were about 1.44 dB LKFS lower in loudness level than the stimuli with $\varphi = 30°$. On the opposite, participants mostly rated the conditions with $\varphi = 270°$ a little bit closer then the stimuli with $\varphi = 30°$, as the stimuli with $\varphi = 270°$ were about 1.15 dB LKFS higher in loudness level than the stimuli with $\varphi = 30°$. These effects of azimuth occurred especially for conditions without normalization.

3 Experiment 2

In Experiment 1, nominal distance had no significant effect on distance ratings for the normalized stimuli except for a single condition, suggesting that participants mostly did not use binaural cues, which vary strongly with nominal distance, for distance estimation. To further verify this somewhat surprising outcome, we conducted a psychophysical two-alternative forced choice (2AFC) test, providing a more sensitive test than the previously used method. In the forced-choice procedure, participants had to judge the relative perceptual distance between two virtual (nearby) sound sources. If at all relevant for distance estimation, barely perceptible differences possibly included in the near-field HRTFs would be easier to detect in such a direct comparison than in the multiple-stimulus test used in Experiment 1. Since the focus of Experiment 2 was on binaural cues, we only examined conditions with loudness normalization and head tracking.

3.1 Method

3.1.1 Participants

Seventy-three participants with an age between 19 and 47 years took part in Experiment 2 (58 males, 15 females, $M = 23.37$ years, $Mdn = 23$ years, $SD = 4.21$). All of them were students in media technology and participated for course credit. None of them had participated in Experiment 1 and they were all naive as to the purpose of this study. Thus, as in the previous experiment, participants had to rely on their life experience in distance perception of nearby sources. Here, only five participants (7%) already had experience with the binaural reproduction system. This small number resulted from the fact that most of the subjects from our commonly used subject pool had already participated in Experiment 1 and thus were not allowed to take part in the second experiment. All participants had self-reported normal hearing.

3.1.2 Setup and stimuli

Head-tracking and loudness normalization were used throughout Experiment 2. In all other aspects, technical setup and stimuli were identical to Experiment 1 (see Sect. 2.1.2).

3.1.3 Procedure

The experiment was a 4 × 3 × 2 within-subjects factorial design with the factors *distance pair* (0.25 m vs. 0.50 m, 0.25 m vs. 0.75 m, 0.25 m vs. 1.00 m, 0.25 m vs. 1.50 m), source *azimuth* (30°, 150°, 270°), and *presentation order* (close – far, far – close). Distances were always compared with reference to the closest distance ($d = 0.25$ m), because the ILD and the spectrum differ most strongly from the

other distances. The factor azimuth describes the azimuthal position of the two virtual sources to be compared, and the factor presentation order describes whether the closest sound source at $d = 0.25$ m was presented first (close – far) or last (far – close).

The procedure of the experiment was as follows. On each trial, a sequence composed of four stimuli was presented. In this sequence, the first and the last two stimuli were always the same, resulting in two stimulus pairs which had to be compared. Thus, the two to-be-compared distances (A and B) were presented twice (A–A–B–B). Each stimulus pair had a total length of 3.50 s (2 × stimulus of 1.50 s + 0.50 s interstimulus interval). Between both stimulus pairs, there was an interstimulus interval of 1.00 s, resulting in a playback time of 8.00 s for each trial. Similar to Experiment 1, we decided to use a rather long test signal as well as stimulus repetitions to provide enough time to move the head during playback.

After playback, participants had to report whether they perceived the second stimulus pair closer or further away than the first one, by pressing the corresponding button on the GUI presented on the tablet. The two buttons were arranged on a vertical line, with the upper one labeled "further away" and the lower one labeled "closer". Participants could neither repeat a trial nor continue without giving an answer. After a response was registered, the next trial followed immediately. A full run consisted of 12 trials per condition, leading to a total of 24 (conditions) × 12 (trials) = 288 trials. The order of conditions was randomized for each participant.

Before starting the test, participants were given instructions about the general procedure. Since most of the participants were new to the field of virtual acoustics, the instruction also included a brief introduction on dynamic binaural synthesis. Furthermore, as in Experiment 1, they were encouraged to perform localization movements with their head if they felt that distance perception improved when doing so. At the same time, because of the different source positions, they were instructed to keep their main line of vision straight ahead and they were not allowed to turn their body, e.g., sideways or to orientate themselves to the sound source. After the instructions, participants conducted a short training session composed of six trials to get familiar with dynamic binaural synthesis and with the test procedure. Altogether, the experiment took about 1 h, including the verbal instruction, the training session, and a short break after half of the trials.

3.1.4 Data analysis

For each subject, the 12 repetitions per condition were averaged first, leading to a quasi-metric variable with a value between 0 and 1 describing the proportion of correct answers. These proportion data follow a binomial distribution, where generally the variance is a function of the mean and variances tend to be small at both ends of the range but large in the middle. As a consequence, it is questionable to use parametric tests with raw proportions, since the assumption of normality and homogeneity of variance might be violated to a certain extent, even though the usual parametric tests like t test or ANOVA are robust to these violations. To lessen this issue, we applied an arcsine square root transformation to the raw data, which is a typical procedure for proportions. The transformation removes the correlation between means and variances and stretches out both ends of the distribution of proportions while compressing the middle, resulting in homogenized variance and improved normality ([52], Chapter 10.2). The statistical analysis, which was quite similar to the one performed in Experiment 1, was conducted using the transformed data.

We analyzed the transformed data using a three-way repeated measures ANOVA with the within-subjects factors distance pair, azimuth, and presentation order. A Jarque-Bera test for normality failed to reject the null hypothesis for 19 out of 24 conditions. With Hochberg [44] correction, the test failed to reject the null for all conditions. We nevertheless corrected for slight violations of ANOVA assumptions using the GG correction [45]. To analyze the data in greater detail, we conduced several nested (GG-corrected) repeated measures ANOVAs as well as one-sample t tests (two-tailed) at a 0.05 significance level. All t tests were corrected for multiple hypothesis testing using the Hochberg [44] method.

3.2 Results

Figure 4 shows the results of the experiment. We separated the data with respect to the factor presentation order, resulting in two subsets, further labeled Close – Far (CF) and Far – Close (FC). For better interpretability, the plots show the raw instead of the transformed data. Most strikingly, Figure 4 (left) shows that all means were below chance level. This suggests that participants did hear a difference between the stimuli, but, surprisingly, perceived the closer source as farther away and vice versa. The corresponding box plots in Figure 4 (right) reveal a rather high variance of the results with whiskers often covering the entire range of proportions, which again shows that different participants interpreted or weighted the available distance cues differently.

The ANOVA summarzied in Table 3 yielded a significant main effect of distance pair and significant interaction effects between distance pair and azimuth as well as between azimuth and presentation order. We further analyzed if the condition means differ significantly from chance by conducting 24 one-sample t tests against arcsine-square root transformed chance level (0.7854), revealing significant deviations from chance for all conditions (all ps $< .001$).

To unpack the observed main and interaction effects (see Tab. 3), we conducted several nested ANOVAs. Overall, the main effect of distance pair was present in both subsets, but appears to be much stronger in subset CF. As regards the main effect of distance, performance decreased with increased distance between the virtual sound sources and consequently with intensified inter-stimulus differences. Paradoxically, this decreased performance can be explained by participants perceiving clearer differences between the stimuli with increased distance between the virtual sound

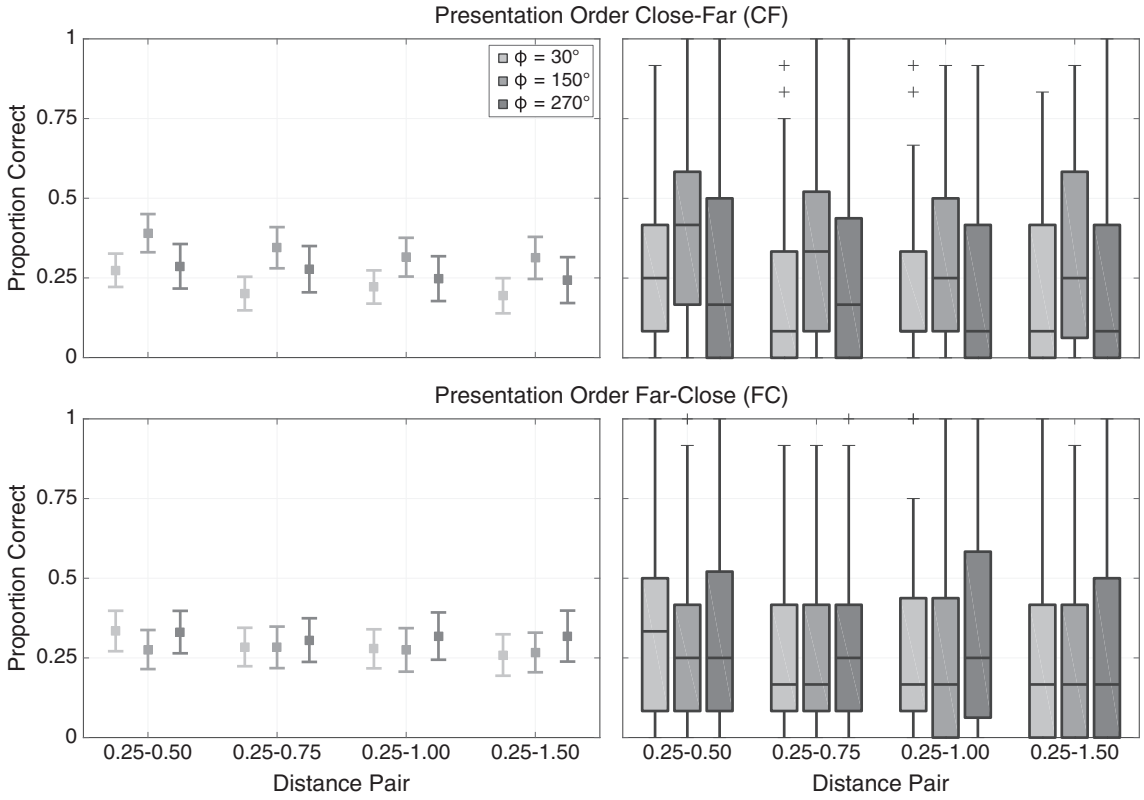


Figure 4. Mean proportions of correct answers (left) and interindividual variation in the proportions (right) as a function of distance pair (abscissa) and nominal source azimuth (shades of gray), separated with respect to presentation order: Close – Far (CF) and Far – Close (FC). For better interpretability, the plots show the raw data instead of the transformed data that was submitted to statistical testing. The error bars in the mean plots (left) display 95% confidence intervals based on the respective one-sample t tests comparing the means of the raw data against non-transformed chance level of 0.5. The box plots (right) show the median and the (across participants) interquartile range (IQR) per condition; whiskers display 1.5 × IQR below the 25th or above the 75th percentile and outliers are indicated by plus signs.

Table 3. Results of the three-way repeated measures ANOVA with the within-subjects factor distance pair (DP), azimuth (Az), and presentation order (PO).

Source	*df*	*F*	*MSE*	ε	η_p^2	*p*
DP	3, 216	21.28	.03	.97	.23	<.001*
Az	2, 144	1.52	.41	.92	.02	.223
PO	1, 72	2.35	.18	1	.03	.130
DP × Az	6, 432	2.31	.03	.94	.03	.036*
DP × PO	3, 216	2.46	.03	.97	.03	.066
Az × PO	2, 144	23.71	.07	.93	.25	<.001*
DP × Az × PO	6, 432	1.53	.03	.95	.02	.169

Note. ε = Greenhouse-Geisser (GG) epsilon, p = GG-corrected p-values. Note that GG correction is appropriate only for within-subject tests, with more than one degree of freedom in the numerator.
*$p < .05$.

sources. The error rate increased because they misinterpreted the provided cues or because wrong cues were given.

Regarding the rather weak distance pair × azimuth interaction effect, two nested ANOVAs indicated a small-but-significant interaction effect between those two factors and a strong main effect of azimuth in subset CF, whereas there was no main effect of azimuth or interaction in subset FC. This can easily be seen in the mean plots, as the values vary clearly as a function of azimuth in Figure 4 (CF – left), but seem to be almost independent of azimuth in Figure 4 (FC – left). These observations also explain the highly significant interaction effect between azimuth and presentation order (Tab. 3). In particular, when the closer source was presented first, the source azimuth had a significant influence on the number of correct answers. Apparently, in this case, the performance for virtual sound sources with an azimuthal position of 30° was much more below chance level than for sources with an azimuthal position of 150°. As opposed to this, the source azimuth had no significant influence on proportions of correct answers if the farther source was presented first.

3.3 Discussion

In line with Experiment 1, participants of Experiment 2 did not correctly employ the strong changes in non-individual binaural cues induced by a variation in nominal sound source distance. Rather, participants predominantly made false responses, which resulted in mean values

significantly below chance level for each tested condition. This indicates that participants did perceive a difference between the stimuli, but made the wrong conclusion with regard to their relative distance. These results might tentatively be explained in two different ways: (a) incorrect use of spectral cues or (b) overcorrection by the normalization method.

The first tentative explanation is based on the signal properties of nearby sound sources: As briefly outlined in Sections 1 and 2.1.3, low frequencies increase relative to high frequencies as a sound source approaches the head, leading to a low-pass filtering character of nearby sound sources. Consequently, stimuli with a distance of 0.25 m were always more dull than stimuli at any other distance. Hence, participants might have most often classified the duller stimulus as the farther source, and the brighter stimulus as the closer source. This assumption is in line with studies from Butler et al. [53] or Little et al. [54], who showed that sounds with decreased high-frequency components relative to low-frequency components are perceived to be further away. Thus, spectrum has a dual role in estimation of distance, as already revealed by Coleman [55], since the relative decrease of high frequency components can be a cue for a source nearby or far away. A dominance of the spectral cue (and the misinterpretation of the provided spectral differences) would also explain the strong effect of distance pair found in the statistical analysis. As outlined above, the proportions decreased as a function of distance pair, which suggests that participants perceived stronger differences with increased distance between the sources. In fact, the spectral differences increase as the distance between the sources becomes larger.

Alternatively, remaining intensity differences between the stimuli may be responsible for the counterintuitive results: As the participants mostly rated the sources at $d = 0.25$ m as further away, it could simply be that the stimuli with a distance of 0.25 m were perceived as slightly quieter than all other stimuli. The perceived loudness differences between the stimuli might have increased as a function of distance pair, which would explain the strong effect of distance pair. Of course, it is also possible that intensity and spectral cues both contributed to the effect. However, if sufficiently strong, the wrong intensity cues most probably masked the spectral cues. As none of the effects was clearly evident in Experiment 1, we assume that differences in spectrum or loudness are more salient in a direct comparison, as provided in Experiment 2, and kind of blur in a multi-stimulus comparison, as in Experiment 1.

Concerning the observed interaction effect between azimuth and presentation order, we could not find any plausible explanation. Thus, it is uncertain why especially the proportions for stimuli with $\varphi = 150°$ were closer to chance level if the closer source was presented first (presentation order close – far). Based on these observations, we can only generally conclude that there were less perceptible differences between these specific stimuli. However, the findings cannot be explained by any signal properties of the stimuli and a detailed exploration of the described effect is beyond the scope of the present study. Indeed, the influence of source azimuth and presentation order on perceived distance of nearby sound sources remains an interesting research question for further studies.

4 Experiment 3

To clarify whether a difference in perceived loudness might explain the surprising below-chance performance in Experiment 2, Experiment 3 directly tested for differences in perceived loudness after normalization. We examined both loudness normalization as employed in this study and amplitude normalization according to Brungart et al. [7] as employed in many previous experiments. To test for perceptible loudness differences between the stimuli, we performed comparison tests according to the SAQI test paradigm [43]. In particular, participants had to judge the relative perceptual loudness difference between two virtual (nearby) sound sources on a bipolar seven-point scale. Similar to Experiment 2, we only examined conditions with (the respective) normalization and head tracking. Regarding the outcome of this follow-up study, we expected to observe that the loudness normalization reduces perceptible loudness differences better than the amplitude normalization.

4.1 Method

4.1.1 Participants

Seventeen male students in media technology or electrical engineering with an age between 19 and 42 years ($M = 24.12$ years, $Mdn = 22$ years, $SD = 5.66$) participated in the experiment on a voluntary basis. Five of them had already participated in Experiment 1, but none of them had taken part in Experiment 2, which was the more recent experiment with largely similar stimuli and procedure. Eight participants (47%) already had experience with the binaural reproduction system and the test environment. All of them reported normal hearing and were naive as to the purpose of this study.

4.1.2 Setup and stimuli

As this was a follow-up study, the experimental setup, the test signal, and the HRTFs were exactly the same as in Experiments 1 and 2 (refer to Sect. 2.1.2). All conditions were with head tracking and the respective normalization method. To get the additional gain values for the amplitude-normalized stimuli, the scaling factor S according to Brungart et al. [7] was calculated for each of the 15 positions (5 distances and 3 directions). This factor is based on the distance of the source from the left and right ears of the listener, such as $S = 1/((50/d_l) + (50/d_r))$, where d_l and d_r is equal to the distance in cm to the left and right ear respectively. The distance between both ears was defined as 0.20 m. The calculated scaling factors were then added to the gain values for the non-normalized stimuli, resulting in amplitude-normalized stimuli when being rendered.

To ensure good comparability between the normalization methods, the gain values were set to the same values at the reference distance of 1.00 m. For stimuli closer or farther away, the gain values obviously differed between the two normalization methods. At the closest distance of 0.25 m, the differences in gain values were greatest. Depending on the direction, the gain values for the amplitude normalization were about 2–4 dB higher than for the loudness normalization in this case.

4.1.3 Procedure

In addition to the independent variables considered in the previous experiment, Experiment 3 involved the two different normalization methods. This resulted in a $4 \times 3 \times 2 \times 2$ within-subjects factorial design (48 conditions) with the factors *distance pair* (0.25 m vs. 0.50 m, 0.25 m vs. 0.75 m, 0.25 m vs. 1.00 m, 0.25 m vs. 1.50 m), source *azimuth* (30°, 150°, 270°), *presentation order* (close – far, far – close), and *normalization method* (loudness, amplitude).

The procedure according to the SAQI assessment for loudness was as follows. On each trial, two different stimuli were presented successively. Corresponding to the stimulus length of 1.50 s and an interstimulus interval of 0.50 s, the total playback time of each trial was 3.50 s. In contrast to Experiment 2, each stimulus was only presented once. However, the total length of the test signal was exactly the same, again to provide enough time for potential head movements and to assure comparability with the other experiments.

After each trial, participants had to rate if they perceived the second stimulus louder or quieter than the first one. The size of the perceived loudness difference had to be given on a bipolar seven-point scale with the comparative scale ends named quieter and louder. The scale was numbered from 0 to 3, with 0 being in the middle and 3 being at both scale ends. It was displayed on the GUI in form of a vertically aligned continuous fader, thus selecting interim values between the given numbers was possible. To avoid a bias towards a specific scale range, the fader knob was reset to 0 (center position) at the beginning of each trial. Thus, if no loudness difference between both stimuli could be perceived, the fader knob could simply remain untouched. By pressing a button displayed on the GUI, participants could continue to the next trial. Each trial was only presented once and participants could not repeat the playback. A full run consisted of 6 trials per condition, resulting in a total of 48 (conditions) $\times$ 6 (trials) = 288 trials. The order of conditions was randomized for each participant.

Before starting the test, participants were given instructions about the general procedure. Participants new to the field of virtual acoustics were also briefly introduced into binaural reproduction technology. Similar to the previous experiments, participants were allowed to turn their head, but they were instructed to keep their front viewing direction and not to turn their body. At the beginning of the test, participants had to conduct a short training session composed of six trials. This way, they could get familiar with the procedure, binaural rendering, and the loudness range of the stimuli. In total, each test session took about one hour, including the verbal instruction, the training trials, and a short break after half of the test.

4.1.4 Data analysis

The statistical analysis was based on normalized mean values per subject. Thus, the 6 repetitions per subject for each condition were averaged first and then normalized to the range from -1 to 1. A Jarque-Bera test for normality failed to reject the null hypothesis for 40 out of 48 conditions. With Hochberg [44] correction, the test failed to reject the null for all conditions. As the ANOVA is very robust to small violations of its assumptions [45], we conducted a GG-corrected four-way repeated measures ANOVA with the within-subjects factors distance pair, azimuth, presentation order, and normalization method. For further analysis, we performed several nested (GG-corrected) repeated measures ANOVAs as well as one-sample t tests (two-tailed) at a 0.05 significance level. To compensate for multiple hypothesis testing, all t tests were corrected using the Hochberg [44] method.

4.2 Results

Figure 5 shows the results pooled over presentation order and separated with respect to normalization method. As can be seen, the participants rated the more distant sources louder than the closer ones if loudness normalization was applied. In contrast, they perceived the more distant sources quieter than the closer ones for the conditions with amplitude normalization. Thus, the plots indicate that both tested normalization methods did not work properly and furthermore led to conflicting results.

The ANOVA summarized in Table 4 yielded a significant main effect of distance pair and normalization method, but no significant main effect of azimuth or presentation order. Furthermore, the analysis revealed a significant interaction effect between azimuth and normalization method, which is of particular interest here, as well as several other two- and three-way interaction effects, which we refrain from discussing in detail in the following in order to focus on the main outcome of the experiment.

To more directly test whether the loudness differences remaining after normalization are significant, we performed 24 one-sample t tests comparing the respective results of the pooled conditions against zero. For 20 conditions, the t tests yielded a significant difference between the respective condition mean and zero ($p < .001$ for all). Only the conditions distance pair 0.25 m vs. 0.75 m, amplitude normalization at all three levels of azimuth as well as the condition distance pair 0.25 m vs. 1.00 m, amplitude normalization, $\varphi = 150°$ were not significantly different from zero.

Furthermore, the pattern of the effect of distance pair differs between the two normalization methods. Whereas the results for loudness normalization have an almost constant offset from zero with only a slight slope as the

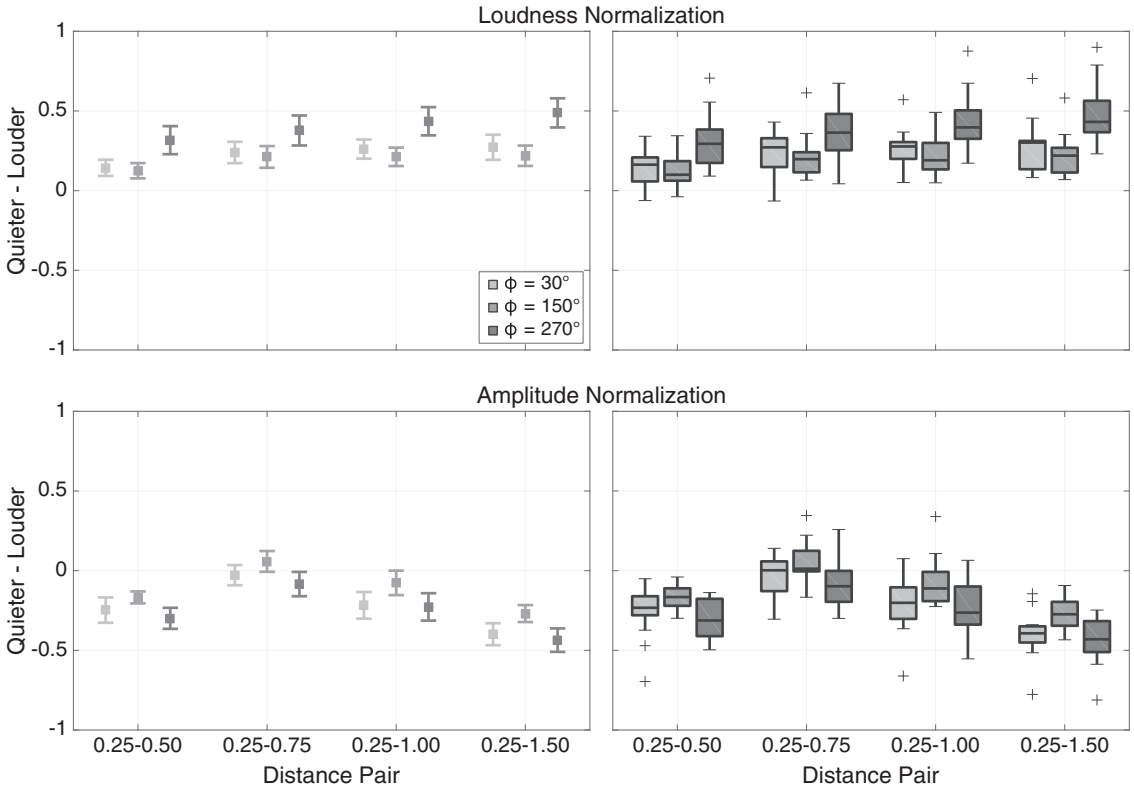


Figure 5. Mean ratings of the SAQI assessment for loudness (left) and interindividual variation in the ratings (right) as a function of distance pair (abscissa) and nominal source azimuth (shades of gray), pooled over presentation order and separated with respect to normalization method. The error bars in the mean plot (left) display 95% confidence intervals based on the respective one-sample t tests comparing the condition means against 0. The box plots (right) show the median and the (across participants) interquartile range (IQR) per condition; whiskers display 1.5 × IQR below the 25th or above the 75th percentile and outliers are indicated by plus signs.

Table 4. Results of the four-way repeated measures ANOVA with the within-subjects factor distance pair (DP), azimuth (Az), presentation order (PO), and normalization method (NM).

Source	*df*	*F*	*MSE*	ε	η_p^2	*p*
DP	3, 48	61.72	.02	.52	.79	<.001*
Az	2, 32	3.03	.11	.99	.16	.063
PO	1, 16	.01	.12	1	0	.919
NM	1, 16	223.47	.21	1	.93	<.001*
DP × Az	6, 96	1.07	.01	.65	.06	.377
DP × PO	3, 48	8.63	.01	.83	.35	<.001*
Az × PO	2, 32	16.27	.05	.62	.50	<.001*
DP × NM	3, 48	80.54	.02	.69	.83	<.001*
Az × NM	2, 32	92.81	.02	.98	.85	<.001*
PO × NM	1, 16	53.45	.04	1	.77	<.001*
DP × Az × PO	6, 96	3.21	.01	.81	.17	.012
DP × Az × NM	6, 96	1.93	.01	.74	.11	.109
DP × PO × NM	3, 48	7.54	.01	.79	.32	.001*
Az × PO × NM	2, 32	6.35	.02	.78	.28	.009*
DP × Az × PO × NM	6, 96	.49	.01	.69	.03	.753

Note. ε = Greenhouse-Geisser (GG) epsilon, p = GG-corrected p-values. Note that GG correction is appropriate only for within-subject tests, with more than one degree of freedom in the numerator.
*$p < .05$.

distance between the sound sources increases, the results for amplitude normalization vary considerably more depending on distance pair and do not to follow a linear trend. Two nested ANOVAs yielded a significant main effect of distance pair with loudness normalization [$F(3, 48) = 42.75$, $p < .001$, $\eta_p^2 = .73$, $\varepsilon = .80$] and with amplitude

normalization [$F(3, 48) = 79.19$, $p < .001$, $\eta_p^2 = .83$, $\varepsilon = .56$]. Thus, intensity cues remain with both normalization methods, but it seems that the offset of the loudness normalization method could be compensated much more easily with a linearly decreasing gain function.

Finally, we examined the influence of azimuth in more detail. The same two nested ANOVAs as described in the previous paragraph showed a significant main effect of azimuth with loudness normalization [$F(2, 32) = 34.23$, $p < .001$, $\eta_p^2 = .68$, $\varepsilon = .81$] and with amplitude normalization [$F(2, 32) = 10.20$, $p < .001$, $\eta_p^2 = .39$, $\varepsilon = .85$]. Eight paired t tests comparing the loudness-normalized conditions with $\varphi = 270°$ against conditions with $\varphi = 30°$ and $\varphi = 150°$ showed that the stimuli with $\varphi = 270°$ led to significantly higher ratings in perceived loudness differences ($p < .01$ for all), i.e., participants perceived the source at $d = 0.25$ m and $\varphi = 270°$ as the least loud.

The results reveal that the loudness normalization attenuates very close sources too much. Thus, sources at $d = 0.25$ m (especially for $\varphi = 270°$) were always quieter than sources farther away, which was particularly perceptible in the direct comparison task. In contrast, the amplitude normalization amplifies very close sources too much and at the same time attenuates the more distant sources. As a result, sources at $d = 0.25$ m were mostly distinctly louder than sources farther away, especially when compared to sources at $d = 1.50$ m. Thus, the amplitude normalization actually results in intensity cues associated with a natural distance shift, and therefore allows for correct distance discrimination based on the intensity cues that it is supposed to remove. Overall, it seems that very close (lateral) sound sources are most critical and that both normalization methods have strong drawbacks.

4.3 Discussion

Experiment 3 revealed that neither the commonly employed amplitude normalization nor the supposedly more suitable loudness normalization correctly removed perceptible loudness differences. Especially for the (lateral) closest source at $d = 0.25$ m, both methods achieved the worst results. Based on these surprising results, we can now explain the counterintuitive results of Experiment 2 as well as some observations of Experiment 1, and can also give a possible explanation for some contradictory findings in the literature.

In Experiment 2, it appears that the inaccurate loudness normalization resulted in intensity cues that participants exploited for distance discrimination. As the reference source at $d = 0.25$ m was always perceived as quieter than any of the other sources, it is not surprising that the participants mostly rated the second (louder) source as closer. Moreover, as shown in Experiment 3, the perceived loudness differences increased as a function of distance pair. This is in line with the findings from Experiment 2 where participants perceived clearer differences between the stimuli with increased nominal distance between the virtual sound sources (main effect of distance pair). Spectral cues, however, most probably played only a minor role or were simply masked by intensity cues. These intensity cues strongly affected the results of the sensitive direct-comparison task employed in Experiments 2 and 3, but apparently played less of a role in the multiple-stimulus comparison procedure in Experiment 1. However, in subset HTNorm of Experiment 1, the normalized source at $d = 0.25$ m and $\varphi = 270°$ was rated further away than the other sources at $d = 0.25$ m. This is in line with the results of Experiment 3, which revealed that this stimulus was perceived as the least loud among all stimuli. Furthermore, the ratings for the loudness normalized stimuli in Experiment 3 follow the same trend (v-shaped pattern) as the ratings for subset HTNorm and $d = 0.25$ m and $d = 0.50$ m in Experiment 1, indicating that participants estimated relative distance in these conditions of Experiment 1 according to the perceived azimuth-dependent loudness differences between the sources. Relatedly, in subset STNorm, stimuli that were nominally farther away were consistently rated as closer, likely because they were perceived as louder. Thus, remaining intensity cues caused by the (erroneous) loudness normalization, as revealed in Experiment 3, are a reasonable explanation for these on first sight surprising findings of Experiment 1, but due to the azimuth effects, a binaural influence on distance estimation cannot be generally ruled out.

5 General discussion

Previous studies have yielded conflicting results regarding the question whether (individual or non-individual) binaural cues contribute to distance perception in the near field. To address this open research question, we conducted three listening experiments using non-individual binaural synthesis. Experiment 1 was designed as a broader study to get a better insight into various potential influences on auditory distance perception. In a multi-stimulus comparison task, subjects had to estimate distance of loudness-normalized and non-normalized nearby sources in static and dynamic binaural synthesis. To isolate binaural cues in the (supposedly) best possible way, we normalized the stimuli in loudness according to ITU-RBS.1770-4. Experiment 2 strictly focused on binaural cues of nearby sound sources and their potential influence on auditory distance perception. Here, subjects had to judge the relative perceived distance between loudness-normalized sources in dynamic binaural rendering. Finally, Experiment 3 assessed the performance of the employed loudness normalization and of the frequently used amplitude normalization proposed by Brungart et al. [7].

The results of Experiment 1 suggest that in most examined conditions, naive listeners did not make use of non-individual binaural cues for distance estimation of nearby loudness-normalized sound sources in anechoic conditions, despite the drastic physical changes in binaural cues (especially in ILDs) due to changes in nominal sound source distance or head movements. In Experiment 2, participants even performed significantly below chance, that is, they mostly interpreted the closer source as the

source farther away. This surprising result was explained by Experiment 3, which revealed that the employed loudness normalization overcorrected so that closer sources were perceived as less loud than farther sources. As a result, the participants in Experiment 2 always compared a slightly quieter source to a somewhat louder source and therefore most probably discriminated distance based on intensity cues, which provided clearly perceptible differences in the sensitive direct-comparison test of Experiment 2. In the multiple-stimulus comparison task of Experiment 1, the effects of remaining intensity cues in conditions with normalized stimuli were weaker, but in line with those observed in Experiments 2 and 3.

Experiment 3 also revealed that previous studies on the effect of binaural cues on distance estimation were likely compromised by the opposite drawback of amplitude normalization. In particular, with amplitude normalization, close sources are still perceived louder than far sources. In other words, the here considered amplitude normalization did not fulfill its intended function. As a consequence, in previous studies employing amplitude normalization, participants might have been able to correctly perceive distance changes based on intensity cues instead of binaural cues. Thus, the present test series clearly demonstrates the problem of normalization as a means to remove intensity cues: with an imperfect normalization, intensity cues remain, which then dominate distance estimation and mask all other cues. Regarding non-individual binaural cues, our results show no clear evidence that, despite their strength in the near field, they contribute to distance estimation of nearby sound sources in anechoic conditions when weak residual intensity cues are still present. However, given the demonstrated drawbacks of the normalization methods causing these residual intensity cues, further studies with other test and normalization methods are needed to clarify the role of binaural cues for distance estimation of nearby sound sources.

Results of our Experiment 1 are in line with Shinn-Cunningham [18] and Kopčo and Shinn-Cunningham [19], who concluded that individual binaural cues are irrelevant for distance perception of nearby sound sources in anechoic conditions [18] and furthermore could not find direct evidence that binaural cues affect distance judgments in reverberant conditions [19]. In a follow-up study using non-individual BRIRs, Kopčo et al. [20] qualified the latter statement by suggesting that the DRR cue is more robust and reliable than the ILD cue, but that the brain actually combines both cues to process distance estimation and does not simply rely on a DRR-to-distance mapping.

In contrast, Brungart et al. [7] as well as Kan et al. [13] and Spagnol et al. [8] for example found that individual as well as non-individual binaural cues affect distance estimation especially for lateral sound sources. In fact, in all these studies, the amplitude normalization proposed by Brungart et al. [7] was applied, which leads to very close sources being presented too loudly according to the results of Experiment 3. Thus, amplitude-normalized stimuli are similar to a natural presentation of sound sources at different distances where closer sources are always (a little) louder. Furthermore, our experiment showed that the perceived loudness differences were particular strong for the lateral sound source ($\varphi = 270°$). Given that intensity is considered as the most dominant cue for distance perception and that even small intensity differences can lead to a change in perceived distance [1, 21], it might be that participants of the above mentioned studies exploited subtle intensity differences between the stimuli for correct distance estimation instead of binaural cues. However, since Brungart et al. [7, 11, 12] roved the level of the normalized stimuli, it is unlikely that participants were able to exploit intensity cues in their studies. Nevertheless, already small intensity differences between the stimuli might be important in localization experiments (without level-roving) as conducted by Kan et al. [13], and especially in a direct-comparison test (relative distance estimation) as for example performed by Spagnol et al. [8], these differences most certainly affect distance estimation. Thus, based on Experiment 3, some results of the above mentioned studies might also be explained by residual intensity cues. However, as previous studies used other stimuli, had other conditions, and applied other individual or generic HRTFs or even used loudspeakers instead of virtual acoustics, the residual-intensity-cue explanation of their results must await further dedicated studies.

Theoretically, roving the level of the stimuli and thus diminishing remaining intensity cues might be a way to compensate for the drawbacks of normalization, especially in studies on absolute distance perception. However, in practice, roving the level seems not expedient for experiments on relative distance estimation applying direct-comparison tasks and normalization. In particular, level-roving would reintroduce intensity cues and thus negate the attempt of the normalization method to eliminate intensity differences between the stimuli. As a result, intensity would most certainly mask any other cue, and listeners would therefore estimate distance purely based on variation in stimulus intensity induced by the roving, i.e., they would most probably perceive the louder source as closer than the quieter source (as confirmed by results of Experiments 2 and 3). Especially naive listeners, such as all participants in the presented listening experiments, seem to be affected by this since they mostly cannot ignore the strong intensity distance cues induced by roving, even if they are explicitly instructed to do so (for direct evidence from a recent pertinent study, see [23]).

Our findings apply to situation where non-individual HRTFs are used. It appears possible that the use of individual HRTFs would affect the results. As discussed in the Introduction (see Sect. 1), it is not clear how exactly and under what circumstances individual HRTFs improve distance perception as compared to non-individual HRTFs. In the special case of distance estimation without DRR cues (anechoic) and without or only weak intensity cues, as tested in the present study, listeners might weight spectral cues more strongly than in natural acoustic conditions. Thus, whereas non-individual spectral cues do not seem to affect distance perception in more realistic listening situations [31], distance perception in our experiments under

anechoic conditions was maybe affected by impaired monaural spectral cues caused by the non-individual HRTFs, similar as for example shown by Baumgartner et al. [36]. Therefore, listeners may have perceived the sources closer (or to some extent less externalized) than would have been the case when using individual HRTFs. However, according to subject reports, listeners perceived sound sources as sufficiently externalized in our experiments. Ultimately, whether individual binaural and spectral cues would have an effect on distance estimation of normalized stimuli is another empirical question which remains to be investigated in future studies.

6 Conclusion

In the present study, we examined how non-individual binaural cues contribute to auditory distance estimation of nearby sound sources. We conducted three experiments in virtual acoustics with naive and untrained listeners. In a multiple-stimulus comparison task (Experiment 1), non-individual binaural cues did not evidently influence distance estimation of nearby sound sources. In a more sensitive direct-comparison task (Experiment 2), listeners might have judged distance based on remaining intensity differences, even with loudness normalization applied. The final experiment (Experiment 3) showed that the loudness normalization applied here, as well as the amplitude normalization introduced by Brungart et al. [7], leave intensity cues that might mask any subtle binaural distance cues. In sum, the present set of experiments has revealed that eliminating all intensity cues without overcorrecting is not a trivial task and that the normalization method should be carefully considered and evaluated when designing a distance perception experiment.

This also means that the influence of binaural cues cannot be correctly investigated with test methods based on stimulus normalization that do not employ additional approaches to effectively suppress remaining intensity cues. Therefore, it remains an open question whether binaural cues contribute to distance estimation of nearby sound sources, and several previous studies cannot be taken as conclusive evidence for or against the role of binaural cues. In fact, our results indicate that several conflicting findings in the literature regarding the role of binaural cues in distance estimation can be explained by differences in normalization methods across the various studies, i.e., maybe subjects only evaluated distance based on remaining salient intensity cues, which masked subtle binaural cues.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF 03FH014IX5-NarDasS). We would like to thank the participants of the experiments for their patience and commitment. We also thank Tim Lübeck for his assistance in conducting the experiments. The research data for this article are available at https://doi.org/10.5281/zenodo.4445283.

Conflict of interest

Authors declared no conflict of interests.

References

1. P. Zahorik, D.S. Brungart, A.W. Bronkhorst: Auditory distance perception in humans: A summary of past and present research. Acta Acustica united with Acustica 91, 3 (2005) 409–420.
2. A.J. Kolarik, B.C.J. Moore, P. Zahorik, S. Cirstea, Shahina Pardhan: Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. Attention Perception & Psychophysics 78, 2 (2016) 373–395.
3. J. Blauert: Spatial Hearing – The Psychophysics of Human Sound Localization. MIT Press, Cambridge, MA, 1996.
4. D.S. Brungart, W.M. Rabinowitz: Auditory localization of nearby sources. Head-related transfer functions. Journal of the Acoustical Society of America 106, 3 (1999) 1465–1479.
5. R.O. Duda, W.L. Martens: Range dependence of the response of a spherical head model. Journal of the Acoustical Society of America 104, 5 (1998) 3048–3058.
6. J.M. Arend, A. Neidhardt, C. Pörschmann: Measurement and perceptual evaluation of a spherical near-field HRTF Set, in Proceedings of the 29th Tonmeistertagung – VDT International Convention, 2016, pp. 356–363.
7. D.S. Brungart, N.I. Durlach, W.M. Rabinowitz: Auditory localization of nearby sources. II. Localization of a broadband source. Journal of the Acoustical Society of America 106, 4 (1999) 1956–1968.
8. S. Spagnol, E. Tavazzi, F. Avanzini: Distance rendering and perception of nearby virtual sound sources with a near-field filter model. Applied Acoustics 115 (2017) 61–73.
9. R.E. Holt, W.R. Thurlow: Subject orientation and judgment of distance of a sound source. Journal of the Acoustical Society of America 46, 6B (1969) 1584.
10. M.B. Gardner: Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space. Journal of the Acoustical Society of America 45, 1 (1969) 47–53.
11. D.S. Brungart: Auditory localization of nearby sources. III. Stimulus effects. Journal of the Acoustical Society of America 106, 6 (1999) 3589–3602.
12. D.S. Brungart, B.D. Simpson: Auditory localization of nearby sources in a virtual audio display, in: Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics, 2001, pp. 107–110.
13. A. Kan, C. Jin, A. Schaik: A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function. Journal of the Acoustical Society of America 125, 4 (2009) 2233–2242.
14. W.E. Simpson, L.D. Stanton: Head movement does not facilitate perception of the distance of a source of sound. Journal of the Acoustical Society of America 86, 1 (1973,) 151–159.
15. L.D. Rosenblum, A. Paige Wuestefeld, K.L. Anderson: Auditory reachability: an affordance approach to the perception of sound source distance. Ecological Psychology 8, 1 (1996) 1–24.
16. B.G. Shinn-Cunningham, S. Santarelli, N. Kopčo: Distance perception of nearby sources in reverberant and anechoic listening conditions: Binaural vs. Monaural Cues, in Poster presented at the 23rd MidWinter meeting of the Association for Research in Otolaryngology, St. Petersburg, Florida, 2000.

17. B.G. Shinn-Cunningham: Distance cues for virtual auditory space, in Proceedings of the First IEEE Pacific-Rim Conference on Multimedia, Sydney, Australia, 2000, pp. 227–230.
18. B.G. Shinn-Cunningham: Localizing sound in rooms, in Proceedings of the ACM SIGGRAPH and EUROGRAPHICS Campfire: Acoustic Rendering for Virtual Environments, Snowbird, Utah, 2001, pp. 17–22.
19. N. Kopčo, B.G. Shinn-Cunningham: Effect of stimulus spectrum on distance perception for nearby sources. Journal of the Acoustical Society of America 130, 3 (2011) 1530–1541.
20. N. Kopčo, S. Huang, J.W. Belliveau, T. Raij, C. Tengshe, J. Ahveninen: Neuronal representations of distance in human auditory cortex. Proceedings of the National Academy of Sciences 109, 27 (2012) 11019–11024.
21. D.H. Ashmead, D. Leroy, R.D. Odom: Perception of the relative distances of nearby sound sources. Perception & Psychophysics 47, 4 (1990) 326–331.
22. G.A. Miller: Sensitivity to changes in the intensity of white noise and its relation to masking and loudness. Journal of the Acoustical Society of America 19, 4 (1947) 609–619.
23. L. Prud'homme, M. Lavandier: Do we need two ears to perceive the distance of a virtual frontal sound source? Journal of the Acoustical Society of America 148, 3 (2020) 1614–1623.
24. ITU-R BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level. International Telecommunications Union, Geneva, 2015.
25. T. Djelani, C. Pörschmann, J. Sahrhage, J. Blauert: An interactive virtual-environment generator for psychoacoustic research II: Collection of head-related impulse responses and evaluation of auditory localization. Acta Acustica united with Acustica 86, 6 (2000) 1046–1053.
26. E.M. Wenzel, M. Arruda, D.J. Kistler, F.L. Wightman: Localization using nonindividualized head-related transfer functions. Journal of the Acoustical Society of America 94, 1 (1993) 111–123.
27. H. Møller, M.F. Sørensen, C.B. Jensen, D. Hammershøi: Binaural technique: do we need individual recordings? Journal of the Audio Engineering Society 44, 6 (1996) 451–469.
28. D.R. Begault, E.M. Wenzel, 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 49, 10 (2001) 904–916.
29. P. Zahorik: Distance localization using nonindividualized head-related transfer functions. Journal of the Acoustical Society of America 108 (2000) 2597.
30. P. Zahorik: Auditory display of sound source distance, in Proceedings of the International Conference on Auditory Displays, 2002, pp. 1–7.
31. V. Best, R. Baumgartner, M. Lavandier, P. Majdak, N. Kop: Sound externalization: A review of recent research. Trends in Hearing 24 (2020) 1–14.
32. G. Yu, L. Wang: Effect of individualized head-related transfer functions on distance perception in virtual reproduction for a nearby source, in Proceedings of the AES International Conference on Spatial Reproduction – Aesthetics and Science, 2018, pp. 1–5.
33. G. Yu, L. Wang: Effect of individualized head-related transfer functions on distance perception in virtual reproduction for a nearby sound source. Archives of Acoustics 44, 2 (2019) 251–258.
34. W.M. Hartmann, A. Wittenberg: On the externalization of sound images. Journal of the Acoustical Society of America 99, 6 (1996) 3678–3688.
35. W. Owenbrimijoin, A.W. Boyd, M.A. Akeroyd: The contribution of head movement to the externalization and internalization of sounds. PLoS One 8, 12 (2013) 1–12.
36. R. Baumgartner, D.K. Reed, B. Tóth, V. Best, P. Majdak, H. Steven Colburn, B. Shinn-Cunningham: Asymmetries in behavioral and neural responses to spectral cues demonstrate the generality of auditory looming bias. Proceedings of the National Academy of Sciences of the United States of America 114, 36 (2017) 9743–9748.
37. A.V. Giner: Scale – conducting psychoacoustic experiments with dynamic binaural synthesis, in Proceedings of the 41st DAGA, 2015, pp. 1128–1130.
38. M. Geier, J. Ahrens, S. Spors: The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods, in Proceedings of the 124th AES Convention, Amsterdam, The Netherlands, 2008, pp. 1–6.
39. C. Pörschmann, J.M. Arend, A. Neidhardt, A spherical near-field HRTF Set for auralization and psychoacoustic research, Proceedings of the 142nd AES Convention, Berlin, Germany, 2017, pp. 1–5.
40. EBU R128: Loudness normalisation and permitted maximum level of audio signals. EBU – European Broadcasting Union, Geneva, 2014.
41. B. Bernschütz: Microphone arrays and sound field decomposition for dynamic binaural recording, Doctoral dissertation, TU Berlin, 2016.
42. C. Pörschmann, C. Störig: Investigations into the velocity and distance perception of moving sound sources. Acta Acustica united with Acustica 95. 4 (2009,) 696–706.
43. A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, S. Weinzierl, A spatial audio quality inventory (SAQI), Acta Acustica united with Acustica 100, 5 (2014) 984–994.
44. Y. Hochberg: A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75, 4 (1988) 800–802.
45. G.V. Glass, P.D. Peckham, J.R. Sanders: Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Review of Educational Research 42, 3 (1972) 237–288.
46. S.W. Greenhouse, S. Geisser: On methods in the analysis of profile data. Psychometrika 24, 2 (1959) 885–891.
47. E.-J. Wagenmakers: A practical solution to the pervasive problems of p values. Psychonomic Bulletin & Review 5 (2007) 779–804.
48. J.N. Rouder, P.L. Speckman, D. Sun, R.D. Morey, G. Iverson: Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review 16, 2 (2009) 225–237.
49. J.N. Rouder, R.D. Morey, P.L. Speckman, J.M. Province, Default Bayes factors for ANOVA designs, Journal of Mathematical Psychology 56, 5 (2012) 356–374.
50. G.R. Loftus, M.E.J. Masson: Using confidence intervals in within-subject designs, Psychonomic Bulletin & Review 1, 4 (1994) 476–490.
51. J. Jarmasz, J.G. Hollands: Confidence intervals in Repeated-Measures Designs: The number of observations principle. Canadian Journal of Experimental Psychology 63, 2 (2009) 124–138.
52. R.R. Sokal, F. James Rohlf: Introduction to Biostatistics, 2nd ed. Dover Publications Inc, Mineola, NY, 2009.
53. R.A. Butler, E.T. Levy, W.D. Neff: Apparent distance of sounds recorded in echoic and anechoic chambers. Journal of Experimental Psychology: Human Perception and Performance 6, 4 (1980) 745–750.
54. A.D. Little, D.H. Mershon, P.H. Cox: Spectral content as a cue to perceived auditory distance. Perception 21, 3 (1992) 405–416.
55. P.D. Coleman: Dual role of frequency spectrum in determination of auditory distance. Journal of the Acoustical Society of America 44, 2 (1968) 631–632.

Appendix

A.1 Overview of discussed studies

Table A.1. Overview of studies investigating the contribution of binaural cues to distance estimation of (nearby) sound sources. (+) Binaural cues contribute. (°) Unclear or mixed findings. (−) Binaural cues do not contribute.

Study	Method	Normalization	Findings and conclusion
Holt and Thurlow [9]	Anechoic conditions. Far-field sources between 1.80 m and 19 m. Participants judged distance in feet.	Level [dB(A)]	(+) Performance improved for lateral sources. Binaural cues are important for distance perception.
Brungart et al. [7]	Anechoic conditions. Near-field sources at distances between 0.15 m and 1.00 m. Participants pointed to the perceived location.	Distance-related amplitude normalization and level-roving	(+) Most accurate distance estimation for lateral sources. ILDs are salient cues for distance estimation.
Brungart and Simpson [12]	Static binaural synthesis with near-field KEMAR HRTFs. Near-field sources at distances between 0.12 m and 1.00 m. Participants pointed to the perceived location.	Signal power and level-roving	(+) Performance worse than in Brungart et al. [7], maybe due to non-individual HRTFs. Still proper distance estimation for lateral sources. ILDs are salient cues for distance estimation.
Gardner [10]	Anechoic conditions. Far-field sources at distances between 0.90 m and 9.00 m. Participants judged distance by choosing a loudspeaker.	Level [dB(B)]	(°) Bad performance for frontal sources. Small head movements led to better performance. Changes in binaural cues might be beneficial.
Kan et al. [13]	Static binaural synthesis with synthesized near-field HRTFs based on individual far-field HRTFs. Near-field sources at distances between 0.10 m and 1.00 m. Participants pointed to the perceived location.	Same as Brungart et al. [7], but without level-roving	(°) Poor performance. Minor distance discrimination for lateral sources at distances < 0.20 m. ILDs are no powerful cues.
Kopčo et al. [20]	Static binaural synthesis with non-individualized near-field BRIRs. Near-field sources at distances between 0.15 m and 1.00 m. 2AFC test – Participants indicated whether the second source was closer or farther than the first one.	Near-ear level [dB (SPL)] and level-roving	(°) Distance estimation based on DRR and ILD cue combination, but DRR cues are more dominant and reliable.
Spagnol et al. [8]	Static binaural synthesis with synthesized near-field HRTFs based on KEMAR far-field HRTFs. Near-field sources at distances between 0.20 m and 1.00 m. 2AFC test – Participants indicated whether the second source was closer or farther than the first one.	Same as Brungart et al. [7], but without level-roving	(°) Poor performance. Similar to Kan et al. [13], slightly improved performance for lateral sources at distances < 0.20 m. ILDs are no powerful cues.
Simpson and Stanton [14]	Quasi-anechoic conditions. Near- and far-field sources at distances between 0.30 m and 2.66 m. Participants rated perceived distance on a scale.	None	(−) No influence of head movements on distance estimation. Binaural cues are not important for distance perception.
Rosenblum et al. [15]	Acoustically normal room. Near-field sources at distances between 0.38 m and 1.10 m. Participants judged the source reachability.	None	(−) No influence of head movements on distance judgment accuracy. Binaural cues are not important for distance judgment.
Shinn-Cunningham et al. [16]	Static binaural synthesis with individual near-field HRTFs/BRIRs. Near-field sources at distances between 0.15 m and 1.00 m. Participants judged distance with a GUI.	No sufficient information	(−) Poor performance. ILD cues do not contribute to distance perception in reverberant conditions and do not provide robust distance percepts even in anechoic conditions.
Kopčo and Shinn-Cunningham [19]	Static binaural synthesis with individual near-field BRIRs. Far- and near-field sources at distances between 0.15 m and 1.70 m. Participants judged distance with a GUI.	Near-ear level [dB (SPL)] and level-roving	(−) Performance was better for lateral sources than for frontal sources and worse without low-frequency energy. In reverberant conditions, only DRR cues are used to judge distance, and not the ILD cues.

Cite this article as:Arend JM, Liesefeld HR & Pörschmann C. 2021. On the influence of non-individual binaural cues and the impact of level normalization on auditory distance estimation of nearby sound sources. Acta Acustica, **5**, 10.

4.3 Measurement and Perceptual Evaluation of a Spherical Near-Field HRTF Set

Arend, J. M., Neidhardt, A., & Pörschmann, C. (2016). In *Proc. of the 29th Tonmeistertagung - VDT International Convention, Cologne, Germany* (pp. 356–363).

Measurement and Perceptual Evaluation of a Spherical Near-Field HRTF Set

(Messung und perzeptive Evaluierung eines sphärischen Satzes von Nahfeld-HRTFs)

Johannes M. Arend[1,2], Annika Neidhardt[3], Christoph Pörschmann[1]
[1] *TH Köln, Institute of Communications Engineering, Cologne, Germany*
[2] *TU Berlin, Audio Communication Group, Berlin, Germany*
[3] *TU Ilmenau, Electronic Media Technology, Ilmenau, Germany*

Abstract

The perceptual refinement of dynamic binaural synthesis has been subject to research for the past years. The basic principle relies on head-related transfer functions (HRTFs), which describe the directional filtering caused by the head, pinna, and torso. However, most systems are based on far-field HRTFs and therefore ignore the acoustical specifics of near-field sound sources. One reason might be that full spherical near-field HRTF sets are rarely available. In this paper, we present an HRTF set of a Neumann KU100 dummy head. The set is freely available for download and contains post-processed impulse responses, captured on a circular and full spherical grid at sound source distances between 0.25 m and 1.50 m. In a subsequent listening experiment using dynamic binaural synthesis, we investigated if the captured binaural cues affect estimated distance of a virtual sound source. The set is useful for various spatial audio applications where nearby virtual sound sources are required, such as auditory displays.

1. Introduction

These days, dynamic binaural synthesis can be regarded as a state-of-the-art approach for headphone-based spatial audio reproduction. The basic principle relies on head-related transfer functions (HRTFs), which describe the directional filtering of the incoming sound caused by the head, pinna, and torso. At this time, a variety of HRTF datasets are available, such as individual HRTF measurements (CIPIC database [1] for example), the established KEMAR dummy head HRTFs [2], or high spatial resolution data of a Neumann KU100 dummy head [3]. The SOFA repository [4] provides an extensive collection of diverse HRTF datasets unified in one data format. In general, the sets are based on anechoic measurements or, in some cases, on simulations. However, most datasets currently available are far-field HRTFs, which means that the sound source used for measurements or simulations was placed at a distance of at least 1 m. Thus, the acoustical specifics of nearby sound sources in the so-called proximal region [5] (the region within 1 m of the listener's head) are simply ignored, even though these features are well known. Stewart [6], Hartley et al. [7], and Brungart et al. [5] for instance theoretically examined the influence of increased head shadowing for nearby sound sources. The studies revealed substantial changes in HRTFs for proximal-region sources. Furthermore, Brungart et al. [5] conducted detailed physical analyses of near-field HRTF data, based on measurements with a KEMAR dummy head. Here, the authors showed a significant increase of interaural level differences (ILDs) as well as an increasing low-pass filtering character of the HRTFs as the sound source approaches the head. Moreover, they outlined the parallax effect for nearby sound sources that especially gains importance when head movements are involved, as is the case with dynamic binaural synthesis. In two subsequent publications, Brungart et al. [8] [9] investigated auditory localization of nearby sound sources. Concerning auditory distance perception in anechoic environments, they conducted a study where subjects had to estimate distance of various level-normalized stimuli, thus loudness-based distance cues were missing. Their results suggested that the specific binaural features found in the HRTFs for nearby sound sources are an important distance cue in the proximal region. As opposed to this, Shinn-Cunningham et al. [10] [11] found in a similar experiment that binaural cues were irrelevant for proximal-region distance perception in anechoic environments. The contrary results show that further investigations in this topic are needed.

Overall, it becomes apparent that the clearly different features of near-field HRTFs should be considered for auralization purposes. Near-field HRTFs for virtual nearby sound sources might improve the plausibility of the virtual auditory scene. Furthermore, proximal-region effects as well as motion-dependent parallax of virtual nearby sound sources could be implemented satisfactorily. Besides, a set of high resolution near-field HRTFs that is publicly available could be used for further experiments regarding auditory localization of nearby sound sources. So far, there are only a few datasets available, whereby some of them can be freely accessed on the Internet [12] [13] [14] and others hardly can be found [15] [16] [17]. However, none of these datasets provide HRTFs with a high SNR over the full audible bandwidth, measured on a full spherical grid with high angular resolution. For use in virtual acoustics, a high-resolution full spherical dataset has several advantages. First of all, it provides a high number of discrete measurement points according to the used spatial sampling grid. Moreover, the dataset can be transformed to the spherical harmonic domain. This allows for spherical harmonic interpolation, which is valid on the entire audible spectrum, given that the measurement resolution is high enough (see [18, Chapter 3.12.4]). As a result, any arbitrary near-field HRTF can be obtained for the respective measured sound source distance. Thus, measuring full spherical datasets at several positions in the proximal region covers a wide range of possible near-field HRTFs.

ISBN 987-3-9812830-7-5

In this paper, we present such a full spherical HRTF database of a Neumann KU100 dummy head, measured with high angular resolution at sound source distances between 0.50 m and 1.50 m. To our knowledge, there is no other full spherical near-field HRTF dataset (of a KU100 dummy head) available so far. Additionally, we captured HRTFs on a circular grid at distances between 0.25 m and 1.50 m, also presented here. The final set is considered to be useful for various auralization applications, like auditory displays or architectural acoustics. Therefore, the focus was on precise positioning, high SNR and full audible bandwidth. Based on the new HRTF set, we conducted several listening experiments. In one experiment, which is presented in this paper, we investigated if the HRTFs can be applied to code distance and if appropriate distance estimation is still possible when natural level differences between the stimuli are missing. The paper is structured as follows. Section 2 describes the HRTF measurements including the measurement setup, the applied post-processing and a technical evaluation of the final HRTF dataset. Section 3 provides the perceptual evaluation of the measured near-field HRTFs. It outlines the used test design as well as preliminary results. Finally, section 4 concludes the paper with a short summary of the measurements and the main findings.

2. HRTF Measurements

2.1. Setup

The HRTF measurements were performed in the anechoic chamber of the acoustics laboratory at TH Köln. The chamber has dimensions of 4.5 m × 11.7 m × 2.30 m (W×D×H) and a low cut-off frequency of about 200 Hz. The sound source was a Geithain RL906 loudspeaker, which has a two-way coaxial design and a flat on-axis magnitude response from 50 Hz to 20 kHz (±3 dB). Thus, the loudspeaker approaches the ideal of an acoustic point source and allows measuring HRTFs in almost the full audible frequency range. The VariSphear measurement system [19] was used for precise positioning of the Neumann KU100 dummy head at the spatial sampling positions and for capturing head-related impulse responses (HRIRs), which is the time-domain equivalent of HRTFs. The impulse responses were measured according to two different spatial sampling grids with two different VariSphear setups: a circular grid where the dummy head was fully rotated in the horizontal plane in steps of 1°, and a Lebedev full spherical grid with 2702 points. The latter is well suited for spherical harmonic interpolation of HRTFs (see [18, Chapter 3.12.4]), which is one possible application of the dataset. Figure 1 shows the respective grids. For the circular grid measurements, the dummy head was mounted on a thin microphone stand, which again was fixed on the rotatable base plate of the VariSphear. When conducting the Lebedev grid measurements, the dummy head was fastened on the robot arm of the VariSphear. In combination with the rotatable base plate, this setup allowed for full 3D rotation of the head on a virtual sphere.
In total, nine HRIR datasets were captured. The circular grid was measured for five sound source distances (0.25 m, 0.5 m, 0.75 m, 1.00 m, 1.50 m) whereas the Lebedev grid was measured for only four distances (0.5 m, 0.75 m, 1.00 m, 1.50 m).

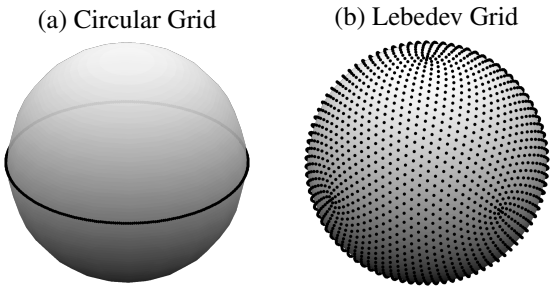


Fig. 1: Measured spatial sampling grids: Circular grid with steps of 1° in the horizontal plane (a) and Lebedev grid with 2702 points (b).

The closest distance was skipped here because the back of the robot arm would have touched the loudspeaker. A cross-line laser was used for precise positioning of the dummy head and the loudspeaker. For both setups, exact alignment of the head was checked for various sampling positions. The distance between the loudspeaker and the entrance of the dummy head's ear canal was determined accurately with a laser distance meter. This procedure was repeated for each new loudspeaker position. The acoustic center of the loudspeaker was always at ear level of the dummy head. Figure 2 exemplary shows the setup for the Lebedev grid measurements at a source distance of 0.5 m. Additionally, omnidirectional impulse responses were captured at the physical origin of the dummy head for all source distances with a Microtech Gefell M296S microphone. These reference measurements provided the basis for the magnitude and phase compensation of the loudspeaker, later on applied in post-processing (see section 2.2).

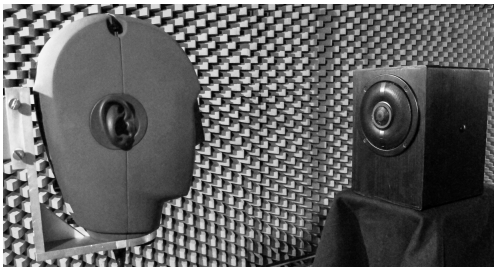


Fig. 2: Measurement setup in the anechoic chamber for the Lebedev grid measurement at a sound source distance of 0.5 m. The sound source was a Geithain RL906 two-way coaxial loudspeaker. The receiver was a Neumann KU100 dummy head, mounted on the VariSphear measurements system.

The excitation signal for all measurements was an emphasized sine sweep with +20 dB low shelf at 100 Hz. With 2^{19} samples at 48 kHz sampling rate, the sweep had a length of about 11 s, which allowed good robustness against background noise. The loudspeaker was driven at about -9 dB below its maximum permissible sound power level and the measurement peak level was always at about -6 dBFS. These settings yielded measurements with an overall SNR of about 90 dB. An RME Fireface UFX audio interface was used as AD / DA converter and microphone preamp. The whole measurement procedure was administered with the VariSphear software. Besides the motor control and impulse response capture modules, the software provided automatic error detection which checked every measured impulse response for noticeable variations with reference to previous measurements. This process ensured validity of all obtained impulse responses.

ISBN 987-3-9812830-7-5

Even though the measurements were conducted with great care, there are several shortcomings, which should be considered. First of all, the loudspeaker might violate the assumption of an acoustic point source in the proximal region (< 1.00 m). Moreover, there might be multiple reflections between loudspeaker and dummy head at close distances, resulting in HRTFs with increased ripple because of interferences. Another serious issue is the influence of the robot arm used for the Lebedev gird measurements. Whereas reflections at the arm are more or less negligible for frontal sound incidence ($\varphi = 0°$, $\delta = 0°$), the arm causes distinct shadowing effects for sound incidence from the rear ($\varphi = 180°$, $\delta = 0°$), which intensify with decreasing sound source distance (see section 2.3 for a more detailed explanation). Thus, the dataset (in particular the Lebedev grid data) should be considered as a valuable set for auralization purposes rather than as a basis for sensitive listening experiments.

2.2. Post-Processing

First, the raw measurement data were carefully truncated, windowed and transformed to the *miro* (measured impulse response object, [3]) format. Working with the MATLAB based miro data type allowed easy access to the datasets and convenient management of further processing. The two major aims of the post-processing were to achieve full range HRTF datasets by extending the low frequency range of the raw measurements and to compensate the influence of the loudspeaker by inverse FIR filtering. Most of the processing is based on the implementation and explanation from Bernschütz [3]. Thus, the following section focuses on the main aspects of the procedure and briefly outlines the processing steps and the technical motivation, whereas Bernschütz provides a more detailed explanation in his publication.

Adaptive Low Frequency Extension The low frequency range of raw HRTFs involves several inaccuracies. First of all, small loudspeakers, which are required for near-field measurements, typically fail to reproduce low frequencies (e.g. below 50 Hz) at adequate sound pressure levels. This leads to HRTFs with a distinct low frequency roll-off. Furthermore, particularly at low frequencies, the loudspeaker induces serious group delay. As a result, the HRIRs are more spread in time and thus more filter taps are required to cover the full audible frequency range. Another great problem is the sound field in the anechoic chamber below its cut-off frequency, where room modes and reflections arise. Because of this modal behavior, raw HRTF measurements show room and position dependent peaks and dips in the lower frequency range and therefore perform poorly when auralizing low frequency content. As a consequence, post-processing HRTFs at low frequencies is mostly necessary when full range datasets are required.

Replacing the low frequency range by an analytic expression is one well-suited approach for low frequency processing of HRTFs [3], [20]. For this purpose, Bernschütz [3] developed an algorithm for adaptive low frequency extension (ALFE), which we used for post-processing. The approach assumes that at frequencies below 400 Hz, pinna and ear canal of the dummy head hardly affect the HRTF and that the head itself has only minor influence on the sound field. According to that, it is reasonable to process HRTFs in order to obtain a flat magnitude response below a certain corner frequency less or equal than 400 Hz. Briefly speaking, the used ALFE-algorithm works as follows. The raw HRTF is high-pass filtered at a certain crossover frequency with a 24 dB/Oct Linkwitz-Riley filter and a matched low frequency extension (LFE) is attached, substituting the original low frequency component. This LFE corresponds to a time shifted Dirac delta function $\delta(n)$, adjusted in level according to the original low frequency component and low-pass filtered with the crossover filter. To match the phase slope of the filtered raw HRTF and the LFE around the crossover frequency, a first-order all-pass filter is applied. Since the algorithm is input-dependent, every raw HRTF as well as the reference measurements were processed separately. The crossover frequency was always set to 200 Hz whereas the cut-off frequency of the all-pass filter had to be adjusted per sound source distance. Figure 3 illustrates the described ALFE-processing in frequency domain. The improved HRTFs show a flat magnitude response below 200 Hz and, when examined in time domain, considerably less group delay.

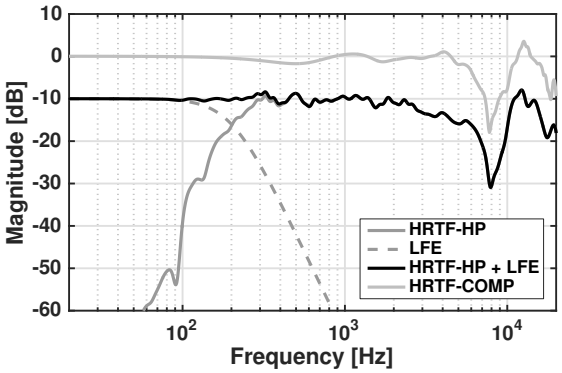


Fig. 3: HRTF post-processing applying adaptive low frequency extension (ALFE) and magnitude / phase compensation. *HRTF-HP*, 1/12-oct. smoothed high-pass filtered raw HRTF (left ear, $\varphi = 0°$, $\delta = 0°$, sound source distance = 1.50 m); *LFE*, low frequency extension - time shifted and low-pass filtered Dirac delta function; *HRTF-HP + LFE*, ALFE-processed HRTF - summed and phase-matched low and high frequency components; *HRTF-COMP*, final HRTF with ALFE-processing and magnitude / phase compensation.

Magnitude and Phase Compensation In a next step, magnitude and phase compensation were applied for further optimization. Therefore, we designed a specific compensation filter for each source distance, based on the ALFE-processed reference measurements. The respective compensation filter was implemented as a Hann-windowed FIR filter, basically describing the appropriately inverted frequency and phase response of the corresponding reference. Filtering all measurements removed further artifacts caused by the loudspeaker, like variations in magnitude response and remaining group delay. As a result of the compensation in time domain, the HRIRs could finally be truncated to 128 taps at 48 kHz sampling rate, while still maintaining the full spectral bandwidth. The length of the head and tail window was set appropriately in the miro files to ensure only negligible influence when windowing is applied. Figure 3 shows an example of a final HRTF in frequency domain.

ISBN 987-3-9812830-7-5

Final Processing In a last processing step, all datasets were slightly leveled so that the HRTFs for sound incidence from the front and from the rear approximate a magnitude of 0 dB at DC. This was more an aesthetic rather than a much-needed step since the deviations from 0 dB at DC were 1 dB at most. The leveling was not applied to the circular dataset with distance of 0.25 m because the peak level of HRIRs for lateral sound incidence would have exceeded 0 dBFS. However, even though all dataset were peak normalized, reconstructing the distance-dependent level differences is still possible based on the normalization factors listed in the miro metadata. Finally, the miro files were converted to the more common SOFA format [4] to provide usability for a wider user group.

2.3. Technical Evaluation

Near-field HRTFs usually have typical signal properties depending on the distance to the sound source, distinguishing them clearly from common far-field HRTFs. Brungart et al. [5, 8, 9], for example, presented a range of such near-field features in their extensive research on nearby sound sources. To check if our new HRTF set also shows the expected characteristics, we examined the final datasets carefully and extracted some of the main features. Moreover, we reviewed all data to check for any deficiencies caused by the measurement setup or post-processing. Please note that all of the following plots showing HRTF properties are based on the circular grid sets, mainly because these sets do not suffer from the influence of the robot arm and because the characteristics are mostly shown in the horizontal plane anyway.

One prominent feature of near-field HRTFs is the increase of ILDs (Interaural Level Differences) as a function of source proximity. According to Brungart [5], especially at sound source distances below 0.5 m, this rise of ILDs is dramatic. Hence, ILDs of near-field HRTFs show the typical increase as the source moves lateral to the head, which is basically caused by (frequency dependent) head shadowing effects. However, since these shadowing effects are much stronger at the contralateral ear and the magnitude at the ipsilateral ear increases simultaneously, the resulting ILDs are distinctly higher [5]. This effect can be easily observed in Figure 5(a), which shows the ILDs of our presented HRTF set for a sound source in the horizontal plane. Whereas the ILDs at the sound source distances 1.50 m, 1.00 m and 0.75 m are more or less similar, they start to increase at a distance of 0.5 m and escalate at the closest distance of 0.25 m. These ILDs up to about 23 dB might provide a relevant cue for distance perception in the proximal region. Off course, ILDs are frequency dependent; a fact also investigated in the context of near-field HRTFs by Brungart et al. [5].

Next, we examined the ITDs (Interaural Time Differences) of the presented HRTF set. Figure 5(b) displays the respective ITDs, calculated by the threshold onset method [21] including 10 times oversampling for more precise onset detection. As expected, the ITDs increase as the source moves lateral to the head and usually peak at about 90° and 270°. Both, the depicted ITDs and ILDs, show the familiar direction-dependent influence of the pinna and the head. However, unlike the ILDs, the ITDs are barely influenced by sound source distance, which is also in line with observations of Brungart et al. [5]. A closer look at Figure 5(b) reveals a slight increase of the time differences as distance decreases, leading to a maximum of about 742 μs at lateral positions and at a sound source distance of 0.25 m. This small rise appears because the length of the path from the ipsilateral to the contralateral side of the head increases as the source approaches. It goes without saying that ITDs and their behavior in the proximal region are also frequency dependent effects, as described more precisely in Brungart et al. [5].

Another prominent effect is the low-pass filtering character of proximal-region sources, meaning that sound sources are getting darker in timbre as they approach the head [5]. This effect is strongest for very close distances and sound sources at the front or rear. It appears because the ears are in the acoustic shadow zone of the head, which mainly damps higher frequencies. The spectral difference between the HRTF at 0.25 m and 1.50 m for frontal sound incidence, shown in Figure 4, demonstrates the described low-pass character. Again, it might be possible that this effect serves as a monaural cue for distance estimation in the proximal region.

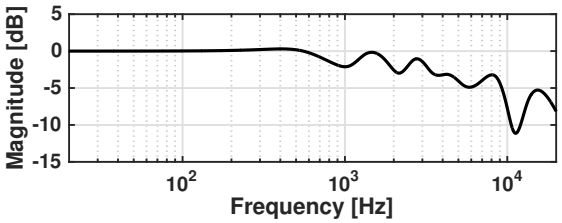


Fig. 4: Spectral difference (1/3-oct. smoothed) between the HRTFs at a source distance of 1.50 m and 0.25 m (left ear, $\varphi = 0°$, $\delta = 0°$). The result illustrates the low-pass filtering effect of proximal-region sources for frontal sound incidence.

In another analysis, we took a closer look at the influence of the robot arm. Therefore, we compared the Lebedev grid with the circular grid measurements at two distances (0.50 m, 1.50 m) and for sound incidence from the front ($\varphi = 0°$, $\delta = 0°$) and from the rear ($\varphi = 180°$, $\delta = 0°$) . By calculating the spectral differences between the respective Lebedev grid and circular grid HRTFs, the influence of the robot arm on the magnitude spectrum can be determined. As depicted in Figure 6, the robot arm only slightly affects the HRTFs for frontal sound incidence. The effect is more or less independent of sound source distance, mainly because the gap between the dummy head and the reflecting robot arm at the back of the head is alway the same. Overall, the reflections at the robot arm cause some minor interference artifacts in the final Lebedev grid HRTFs, starting at about 700 Hz. In the frequency range between 700 Hz and 20 kHz and at the distance of 0.50 m, the ripple has a mean of about 0.55 dB (SD = 0.59 dB) and a maximum absolute value of 2.25 dB at 7.2 kHz. For this particular case, the perceptual influence of the artifacts might be relatively small. For sound incidence from the rear, however, the robot arm causes strong shadowing effects and interferences, as shown in Figure 6. Here, at the distance of 0.50 m, the spectral difference has a mean of about 6.51 dB (SD = 4.61 dB) and maximal damping values of about 10 - 15 dB at frequencies above 10 kHz. At 1.50 m, especially the high frequency damping effect above 10 kHz is weaker, basically because the robot arm does not cover the tweeter

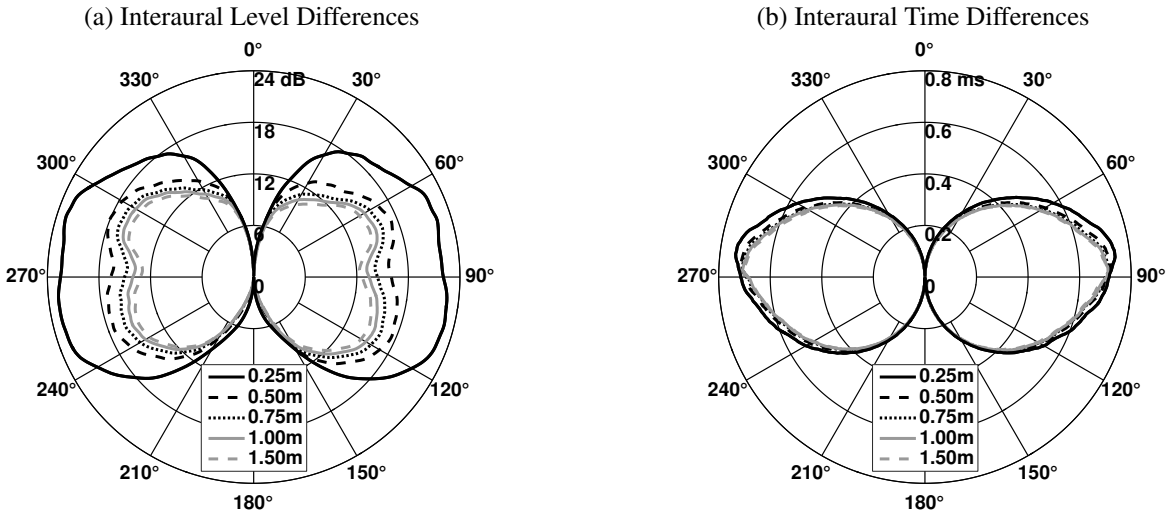


Fig. 5: Interaural Level Differences (a) and Interaural Time Differences (b) of the presented HRTF dataset. The angle represents the azimuth of the sound source (φ). The radius describes the magnitude of the level differences (in dB) or time differences (in ms).

of the loudspeaker. Nevertheless, these Lebedev grid HRTFs clearly suffer from the influence of the robot arm, regardless of the source distance. Both HRTFs lack high frequencies, which is plainly audible in auralizations, especially when compared to the corresponding circular grid HRTF.

Moreover, our signal analysis showed that the reflections between the loudspeaker and the dummy head only affect the post-processed circular grid HRTFs for a distance of 0.25 m. Truncating the HRIRs to 128 taps removed the reflections in the datasets for higher distances, simply because their delay exceeds the length of the HRIRs. Nevertheless, as already mentioned in the paragraph about shortcomings of the measurements in section 2.1, using the HRTF set for sensitive listening experiments should be carefully considered. However, regarding the key features of the presented HRTF set (thoroughly post-processed full range HRTFs, several distances in the near and far field, circular and full spherical grid), it is well suited for many auralization applications.

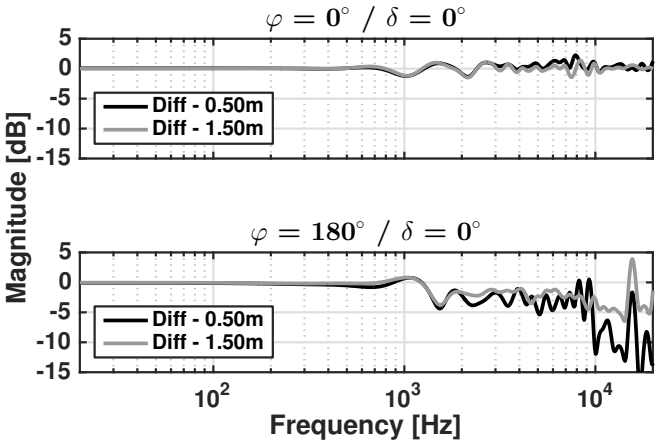


Fig. 6: Spectral differences (1/12-oct. smoothed) between circular grid and Lebedev grid HRTFs at a source distance of 0.50 m and 1.50 m and for sound incidence from the front ($\varphi = 0^\circ$, $\delta = 0^\circ$) and rear ($\varphi = 180^\circ$, $\delta = 0^\circ$). The results illustrate the minor influence of the robot arm on the Lebedev grid HRTFs for frontal sound incidence and its strong effect on the Lebedev grid HRTFs for sound incidence from the rear. In the latter case, the Lebedev grid HRTFs clearly lack high frequencies in comparison to the circular grid HRTFs.

3. Perceptual Evaluation

Based on the new HRTF set, we conducted several listening experiments within the context of auditory distance perception. In this paper, we present preliminary results for one part of this test series. Here, the basic task was to estimate auditory distance to a virtual sound source in dynamic binaural synthesis. The presented study served to investigate if the HRTFs can be applied to code distance and if appropriate distance estimation is still possible when natural level differences between the stimuli are missing. The latter is of particular interest, since the significant changes of binaural and monaural cues for a sound source in the near-field suggest that it is possible to distinguish distance (in the proximal region) even without the prominent factor level difference. Please note that the subjects had no previous training in distance estimation of nearby sound sources. Thus, they had to rely on their life experience in perceiving near-field sound sources.

3.1. Method

Participants Two females and 13 males aged between 21 and 28 years (M = 24.1 years, SD = 2.23) participated at this stage of the experiment. Most of them were students in media technology or electrical engineering. Thirteen participants already took part in previous listening experiments and thus were familiar with the binaural system. None of the subjects reported any hearing problems.

Setup The experiment took place in the anechoic chamber at TH Köln, which ensured a low background noise level of less than 20 dB(A). The experiment was implemented, controlled, and executed with the MATLAB-based software Scale [22], which also accessed the SoundScape Renderer [23] for binaural rendering. To acquire horizontal head movements, a Polhemus Fastrak head tracking system was used. Vertical or translational head movements were disregarded. The subjects entered their responses on a tablet computer (iPad). The audio signal was presented over AKG K-601 headphones. Headphone compensation was applied according to [3] in order to equalize the binaural chain.

Materials The anechoic test signal was a pink noise burst sequence with a burst length of 1500 ms (including 10 ms cosine-squared onset/offset ramps) and pauses of 500 ms. For the listening experiment, we used the circular grid measurements for all five distances from 0.25 m to 1.50 m. Per distance, we tested for three different sound incidence angles (φ = 30°, 150° and 270°). As already mentioned, we also wanted to test if appropriate distance estimation is still possible without natural level differences. Therefore, we prepared a second set of HRTFs, loudness-normalized with regard to the pink noise test stimuli according to ITU-R BS.1770. The playback level for the loudness-normalized conditions was at about 61 dB(A) Leq. For the non-normalized conditions, we assigned this playback level to a sound source distance of 1 m, resulting in a maximum playback level of about 79 dB(A) Leq for the closest distance of 0.25 m (φ = 270°).

Procedure As already mentioned above, there was no training session and no scale anchoring process. Informal pretests showed that training involved strong learning effects, especially for the normalized conditions: First, test persons could not immediately distinguish between distances, but when they were given feedback, they learned to differentiate based on spectral changes, varying ILDs and head movement. However, we wanted to know if distance perception in the near field works instantaneously without prior knowledge about the auditory scene. Therefore, we only gave a basic instruction about the general procedure and the rating scale.
The listening test was composed of two sessions. In the first session, subjects had to rate the normalized conditions, in the second session the non-normalized ones. Thus, the *normalization* order was blocked across participants. In each session, every participant had to rate the five measured *distances* (0.25 m, 0.50 m, 0.75 m, 1.00 m, 1.50 m) for three different source *azimuths* (φ = 30°, 150°, 270°). This resulted in a 5 × 3 × 2 within-subjects design.
Participants had to rate distance on a seven-point category scale ("very close", "close", "rather close", "medium", "rather distant", "distant", "very distant"); a scale that had been successfully used in earlier experiments [24]. It was allowed to rate interim values between the given categories. The procedure was as follows. For each trial, a user interface was displayed on the tablet computer containing five value faders ranging from "very close" to "very distant" (see Figure 7). The five faders corresponded to the five actual measured distances, thus the subjects had to rate multiple stimuli per trial. The source azimuth was the same for all distances (or faders) within a trial. By touching the respective fader, the participants were able to switch between the corresponding stimuli as often as required. Technically speaking, the HRTF filter-set switched when touching the fader while the noise sequence was played in a loop. The order of the faders per trial as well as the order of the trials itself were randomized. The procedure was repeated 10 times per azimuth, thus a full run consisted of 30 trials (with five distance ratings per trial). The listeners were encouraged to move their head during the estimation process in the form of (small) localization movements. However, they had to keep their front viewing direction because of the different source directions. In total, the test lasted for about one hour including the verbal instruction, one short break, and three post-experiment questions.

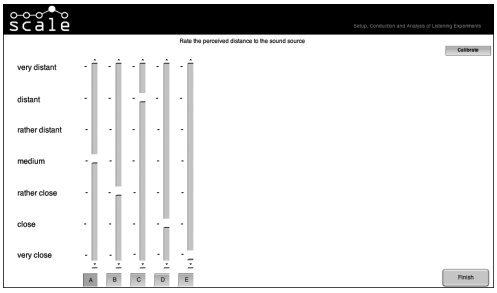


Fig. 7: User interface of the experiment. The left side displays the seven-point category scale. The five faders correspond to the five actual measured distances, randomly ordered for each trial.

3.2. Results

The following statistical analysis is based on the mean value per subject, thus the 10 trials per subject for each condition were averaged first. A 5 × 3 × 2 repeated measures ANOVA (distance, azimuth, normalization) with Greenhouse-Geisser (GG) correction [25] (for tests with more than one degree of freedom in the numerator, where GG is appropriate) was conducted. The ANOVA yielded a significant distance main effect ($F(4,56) = 71.91$, $p < .001$, $\eta_p^2 = .84$, $\epsilon = .41$) as well as significant interaction effects of distance × azimuth ($F(8,112) = 6.97$, $p = .004$, $\eta_p^2 = .33$, $\epsilon = .24$), distance × normalization ($F(4,56) = 34.78$, $p < .001$, $\eta_p^2 = .71$, $\epsilon = .30$) and distance × azimuth × normalization ($F(8,112) = 9.65$, $p = .001$, $\eta_p^2 = .41$, $\epsilon = .26$). Figure 8 presents the respective means of estimated distance per normalized (a) and non-normalized (b) conditions, averaged over subjects. The error bars display 95% within-subject confidence intervals [26], based on the error term of the distance main effect. The interaction effect of distance and azimuth is mainly caused by the variances for conditions with a source distance of 0.25 m (see Figure 8(a)). A repeated measures ANOVA without these conditions confirmed this: here, the interaction effect of distance and azimuth was not significant anymore ($p = .05$). More interesting seems to be the interaction effect of distance and normalization, which is why the mean plots in Figure 8 are split relative to the factor normalization. Presenting the results this way suggests that participants failed to distinguish distances for the normalized conditions (see Figure 8(a)). Without loudness-normalization, meaning with natural level differences between the stimuli, the results are as expected: the subjects rated according to the actual measured distances (see Figure 8(b)). A nested repeated measures ANOVA, each for the normalized and the non-normalized conditions, supported this assumption. Whereas there was no significant main effect of distance ($p = .71$) or azimuth ($p = .76$) for the normalized conditions, there was a strong distance main effect ($F(4,56) = 135.73$, $p < .001$, $\eta_p^2 = .91$, $\epsilon = .36$) for the non-normalized ones. These results indicate that the binaural and monaural cues characterizing sources in the near- and far-field do not influence distance estimation, even though the actual signal variations are huge in some cases (see section 2.3). The results were quite surprising, especially because we expected an effect of these cues, similar to Brungart et al. [8]. However, our results are rather in line with the findings from Shinn-Cunningham et al. [10] [11]. Hence, it appears that the participants rated

ISBN 987-3-9812830-7-5

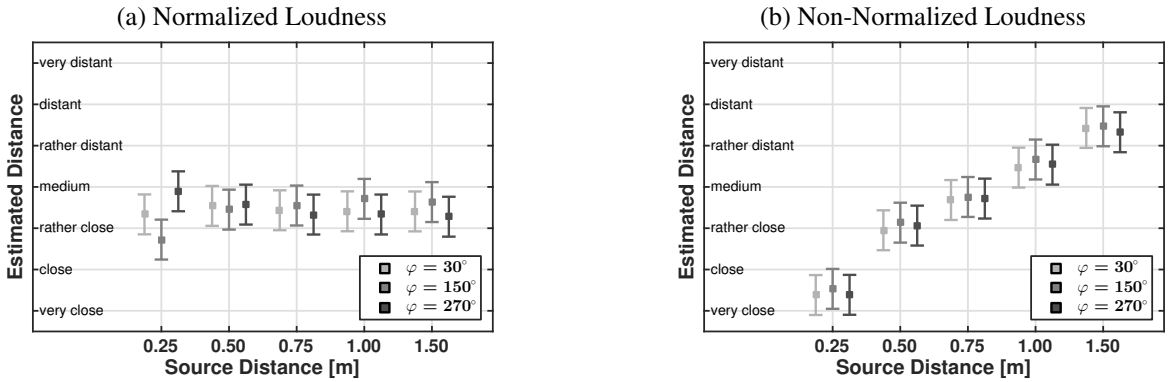


Fig. 8: Mean estimated distances for loudness-normalized (a) and non-normalized (b) conditions as a function of source distance (abscissa) and source azimuth (colors). The error bars denote 95% within-subject confidence intervals based on the respective main effect of distance.

distance mainly based on signal amplitude. Furthermore, a closer look at the results for normalized conditions with a distance of 0.25 m revealed large inter-subject differences. Some subjects correctly rated the proximal-region sources with their low-pass filtering character as close to the head (see section 2.3), whereas others assigned these conditions to very large distances, most likely because they interpreted the muffled sound as a result of high frequency energy dissipation. Overall, most subjects seemed not to have much experience in perception of nearby sound sources. Regarding the normalized conditions, the participants mostly stated that distance estimation was rather difficult and that they were very uncertain about the correct order of the stimuli.

4. Conclusion

Proper auralization of nearby sound sources requires near-field HRTFs with their specific features. In this paper, we presented a near-field HRTF set of a Neumann KU100 dummy head. The set contains post-processed impulse responses, measured according to two different spatial sampling grids: a Lebedev full spherical grid with 2702 points at four sound source distances (0.5 m, 0.75 m, 1.00 m, 1.50 m) and a circular grid with steps of 1° in the horizontal plane at five distances (0.25 m, 0.5 m, 0.75 m, 1.00 m, 1.50 m). After detailed explanations of the measurement setup and of the applied post-processing, we presented a technical evaluation of the final HRTF set and showed the typical (and expected) features of the near-field HRTFs. The final set served as the basis for a series of listening experiments within the context of auditory distance perception in anechoic environments. In the study presented in this paper, we investigated if the HRTFs can be applied to code distance and if appropriate distance estimation is still possible when natural level differences are missing. As expected, the preliminary results showed that distances can be distinguished when loudness-based distance cues exist, thus when the stimuli are not normalized in loudness. However, we observed that subjects could not estimate distances for loudness-normalized stimuli. These findings suggest that binaural cues do not affect distance estimation and vice versa, that auditory distance perception in anechoic environments mainly depends on loudness-based distance cues.

To go further into this issue, additional listening experiments need to be done. As already mentioned, the presented study is part of a larger test series concerning auditory distance perception of nearby sound sources. In an ongoing study, we investigate the influence of head tracking on distance estimation. Furthermore, it would be interesting to examine if a preceding training session influences distance estimation of nearby sound sources.

Apart from the listening experiments, which focus on specific research questions, our primary intention was to provide a freely available near-field HRTF dataset which is well suited for auralization purposes. Therefore, the set is available in the miro and SOFA format under a Creative Commons CC BY-SA 4.0 license and can be downloaded at: *http://audiogroup.web.th-koeln.de/ku100nfhrir.html*.

5. Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (BMBF) under the support code 03FH014IX5-NarDasS. The authors thank all participants of the listening experiment. We thank Philipp Stade, Tim Lübeck and Patrick Pereira for their support during the experiments. The authors wish to thank Benjamin Bernschütz for his advice concerning the measurements and the post-processing.

6. References

[1] Algazi, V. R., Duda, R. O., and Thompson, D. M., "The CIPIC HRTF Database," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 99–102, 2001.

[2] Gardner, W. G. and Keith, D. M., "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.*, 97(6), pp. 3907–3908, 1995.

[3] Bernschütz, B., "A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100," in *Proceedings of the 39th DAGA*, pp. 592–595, 2013.

[4] Majdak, P., Iwaya, Y., Carpentier, T., Nicol, R., Parmentier, M., Roginska, A., Suzuki, Y., Watanabe, K., Wierstorf, H., Ziegelwanger, H., and Noisternig, M., "Spatially Oriented Format for Acoustics: A Data Exchange Format Representing Head-Related Transfer Functions," in *Proceedings of the 134th AES Convention, Rome, Italy*, pp. 1–11, 2013.

[5] Brungart, D. S. and Rabinowitz, W. M., "Auditory localization of nearby sources. Head-related transfer functions," *J. Acoust. Soc. Am.*, 106(3), pp. 1465–1479, 1999.

[6] Stewart, G. W., "The Acoustic Shadow of a Rigid Sphere, with Certain Applications in Architectural Acoustics and Audition," *Phys. Rev.*, 33(6), pp. 467–479, 1911.

[7] Hartley, R. V. L. and Fry, T. C., "The Binaural Location of Pure Tones," *Phys. Rev.*, 18(6), pp. 431–442, 1921.

[8] Brungart, D. S., Durlach, N. I., and Rabinowitz, W. M., "Auditory localization of nearby sources. II. Localization of a broadband source," *J. Acoust. Soc. Am.*, 106(4), pp. 1956–1968, 1999.

[9] Brungart, D. S., "Auditory localization of nearby sources. III. Stimulus effects," *J. Acoust. Soc. Am.*, 106(6), pp. 3589–3602, 1999.

[10] Shinn-Cunningham, B. G., Santarelli, S., and Kopco, N., "Distance Perception of Nearby Sources in Reverberant and Anechoic Listening Conditions: Binaural vs. Monaural Cues," in *Poster presented at the 23rd MidWinter meeting of the Association for Research in Otolaryngology, St. Petersburg, Florida*, 2000.

[11] Shinn-Cunningham, B. G., "Localizing Sound in Rooms," in *Proceedings of the ACM SIGGRAPH and EUROGRAPHICS Campfire: Acoustic Rendering for Virtual Environments, Snowbird, Utah*, pp. 17–22, 2001.

[12] Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., and Kollmeier, B., "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, 2009, pp. 1–10, 2009.

[13] Wierstorf, H., Geier, M., Raake, A., and Spors, S., "A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances," in *Proceedings of the 130th AES Convention, London, UK*, pp. 1–4, 2011.

[14] Xie, B., Zhong, X., Yu, G., Guan, S., Rao, D., Liang, Z., and Zhang, C., "Report on Research Projects on Head-Related Transfer Functions and Virtual Auditory Displays in China," *J. Audio Eng. Soc.*, 61(5), pp. 314–326, 2013.

[15] Nishino, T., Hosoe, S., Takeda, K., and Itakura, F., "Measurement of the head related transfer function using the spark noise," in *Proceedings of 18th International Congress on Acoustics*, pp. 1437–1438, 2004.

[16] Hosoe, S., Nishino, T., Itou, K., and Takeda, K., "Development of Micro-Dodecahedral Loudspeaker for Measuring Head-Related Transfer Functions in The Proximal Region," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pp. 329–332, 2006.

[17] Qu, T., Xiao, Z., Gong, M., Huang, Y., Li, X., and Wu, X., "Distance-Dependent Head-Related Transfer Functions Measured With High Spatial Resolution Using a Spark Gap," in *IEEE Transactions on Audio, Speech and Language Processing*, volume 17, pp. 1124–1132, 2009.

[18] Bernschütz, B., *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording*, Doctoral dissertation, TU Berlin, 2016.

[19] Bernschütz, B., Pörschmann, C., Spors, S., and Weinzierl, S., "Entwurf und Aufbau eines variablen sphärischen Mikrofonarrays für Forschungsanwendungen in Raumakustik und Virtual Audio (Design and Construction of a Variable Spherical Microphone Array for Research in Room Acoustics and Virtual Audio)," in *Proceedings of the 36th DAGA*, pp. 717–718, 2010.

[20] Xie, B., "On the low frequency characteristics of head-related transfer function," *Chinese Journal of Acoustics*, 28(2), pp. 116–128, 2009.

[21] Katz, B. F. G. and Noisternig, M., "A comparative study of interaural time delay estimation methods." *J. Acoust. Soc. Am.*, 135(6), pp. 3530–3540, 2014.

[22] Vazquez Giner, A., "Scale - Conducting Psychoacoustic Experiments with Dynamic Binaural Synthesis," in *Proceedings of the 41st DAGA*, pp. 1128–1130, 2015.

[23] Geier, M., Ahrens, J., and Spors, S., "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods," in *Proceedings of the 124th AES Convention, Amsterdam, The Netherlands*, pp. 1–6, 2008.

[24] Pörschmann, C. and Störig, C., "Investigations Into the Velocity and Distance Perception of Moving Sound Sources," *Acta Acustica united with Acustica*, 95(4), pp. 696–706, 2009.

[25] Greenhouse, S. W. and Geisser, S., "On Methods in the Analysis of Profile Data," *Psychometrika*, 24(2), pp. 885–891, 1959.

[26] Loftus, G. R. and Masson, M. E. J., "Using confidence intervals in within-subject designs," *Psychon. Bulletin & Review*, 1(4), pp. 476–490, 1994.

5 Binaural Rendering of Spherical Microphone Array Data

5.1 Binaural Reproduction of Dummy Head and Spherical Microphone Array Data – A Perceptual Study on the Minimum Required Spatial Resolution

Lübeck, T., Arend, J. M., & Pörschmann, C. (2022). *J. Acoust. Soc. Am.*, *151*(1), 467–483. https://doi.org/10.1121/10.0009277

JASA ARTICLE

Binaural reproduction of dummy head and spherical microphone array data—A perceptual study on the minimum required spatial resolution

Tim Lübeck,[a),b)] Johannes M. Arend,[a),c)] and Christoph Pörschmann[d)]
Technische Hochschule Köln—University of Applied Sciences, Institute of Communications Engineering, Cologne, Germany

ABSTRACT:
Dynamic binaural synthesis requires binaural room impulse responses (BRIRs) for each head orientation of the listener. Such BRIRs can either be measured with a dummy head or calculated from the spherical microphone array (SMA) data. Because the dense dummy head measurements require enormous effort, alternatively sparse measurements can be performed and then interpolated in the spherical harmonics domain. The real-world SMAs, on the other hand, have a limited number of microphones, resulting in spatial undersampling artifacts. For both of the methods, the spatial order N of the underlying sampling grid influences the reproduction quality. This paper presents two listening experiments to determine the minimum spatial order for the direct sound, early reflections, and reverberation of the dummy head or SMA measurements required to generate the horizontally head-tracked binaural synthesis perceptually indistinguishable from a high-resolution reference. The results indicate that for direct sound, $N = 9$–13 is required for the dummy head BRIRs, but significantly higher orders of $N = 17$–20 are required for the SMA BRIRs. Furthermore, significantly lower orders are required for the late parts with $N = 4$–5 for the early reflections and reverberation of the dummy head BRIRs but $N = 12$–13 for the early reflections and $N = 6$–9 for the reverberation of the SMA BRIRs.
 https://doi.org/10.1121/10.0009277
(Received 21 June 2021; revised 4 December 2021; accepted 17 December 2021; published online 27 January 2022)
[Editor: Efren Fernandez-Grande] Pages: 467–483

I. INTRODUCTION

The binaural auralization of virtual acoustic environments can be achieved by convolution with binaural room impulse responses (BRIRs). Such BRIRs can either be obtained with impulse response measurements using a dummy head (Stade *et al.*, 2012), calculated from spherical microphone array (SMA) captures (Bernschütz, 2016), or generated by parametric synthesis (Mccormack *et al.*, 2020; Merimaa and Pulkki, 2004; Pulkki, 2007; Tervo *et al.*, 2013) or simulation (Brinkmann *et al.*, 2019; Savioja and Svensson, 2015; Vorländer, 2008). The auralization with the BRIRs directly measured with a dummy head can still be regarded as the ground truth (Brinkmann *et al.*, 2014; Lindau, 2014). Many studies, either on SMA auralization, room simulation, or parametric synthesis, compare to a reference measured with a dummy head (Ahrens, 2019; Ahrens and Andersson, 2019; Bernschütz, 2016; Garí *et al.*, 2019). Ideally, the BRIRs calculated from the SMA captures are equivalent to these dummy head BRIRs, which is why we focus on these two methods: the BRIRs based on the dummy head measurements and the BRIRs synthesized based on the SMA measurements together with a set of head-related transfer functions (HRTFs) of the same dummy head.

The dynamic binaural synthesis, where the sound field is adapted to the listeners' orientation, requires BRIRs for the arbitrary directions. Lindau *et al.* (2008) showed that a grid resolution of 2° in the horizontal and vertical directions ensures artifact-free auralization. Because the dense full-spherical dummy head BRIR sets are costly in terms of measurement effort and memory consumption, often interpolation of the sparse BRIR sets to the desired directions is applied, which introduces artifacts that are possibly degrading the auralization quality.

The impulse response measurements with the SMAs and a set of anechoic HRTFs are an alternative to the full-spherical dummy head measurements. Once the sound field is captured with the SMA, BRIRs for the arbitrary head orientations can be synthesized. These impulse responses can be measured simultaneously with the real-world SMAs or sequentially with the single-microphone measurements using automated systems such as the VariSphear (Bernschütz *et al.*, 2010). The binaural synthesis from the SMA captures also has the advantages that individual HRTFs can easily be integrated and real-time applications can be implemented. For the real-world SMAs, the major limitation is the number of microphone capsules on the array surface, which leads to undersampling errors and impairments of the binaural signals (Lübeck *et al.*, 2020a). Currently, commercially available

[a)]Also at: Technical University of Berlin, Audio Communication Group, Berlin, Germany.
[b)]Electronic mail: tim.luebeck@th-koeln.de, ORCID: 0000-0003-2870-095X.
[c)]ORCID: 0000-0002-5403-4076.
[d)]ORCID: 0000-0003-0794-0444.

SMAs have between 4 and 64 microphones. The sequential SMA measurements on a dense grid, on the other hand, are very time-consuming, which is similar to the dummy head measurements.

Thus, the spatial interpolation of sparsely measured BRIRs, as well as the calculation of BRIRs from the SMAs with a limited number of microphones, introduce audible artifacts in the binaural signals. Hence, the number of spatial sampling points, whose density and arrangement can be specified by a spatial sampling grid of a certain (spatial) order N_{grid}, has a significant influence on the binaural synthesis. This spatial order N_{grid} is strongly related to the spatial resolution, which is why both terms are used interchangeably in this paper. The influence of the spatial order on the binaural synthesis has been investigated in several studies. The listening experiments by Pike (2019, Chap. A.8) showed that the auralization of the HRTFs interpolated in the spherical harmonics (SH) domain up to an order of 35 were indistinguishable to the auralizations of the HRTFs measured at that position. Similar thresholds were found by Arend *et al.* (2021). The studies by Ahrens and Andersson (2019) and Bernschütz (2016) showed that the perceptual differences of the dummy head auralizations and binaural renderings of the SMA data significantly decrease above the SH orders of 7–8. However, so far, no study systematically compared the perceptual influence of the spatial order of the measurement grid of the dummy head and SMA captures on the binaural synthesis. With this work, we intend to further contribute to the understanding of the different influences of the spatial order of the dummy head and SMA auralizations. To comparably scale the spatial resolution of both auralizations, we applied the interpolation of the dummy head measurements in the SH domain and also performed the binaural rendering of the SMA data based on the SH representation of the sound field. Thus, the signal processing of both methods is affected by the same artifacts as a result of the order-limited SH processing but differ in the spatial aliasing effects. Thus, as a major hypothesis, we assume that the SMA renderings require higher spatial orders than the dummy head SH interpolation, which is elaborated in more detail in Sec. II.

The BRIRs usually describe the direct sound incidence, early reflections, and diffuse reverberation, which all contribute to the spatial auditory perception of the room acoustics in different ways (Kuttruff, 1973, Chap. 4.2). Humans mainly use the direct sound for the sound source localization. It is followed by a number of early reflections, which evoke other perceptual effects, e.g., the apparent source width, perceived distance, timbre, and spaciousness (Barron, 1971; Olive and Toole, 1988). The diffuse sound field is defined by the uniform sound pressure and incident intensity distribution (Jeong, 2016). Thus, in an ideally diffuse sound field, reverberation has no perceivable directional component but contributes to the perception of other room acoustical features, i.e., the perceived room size or listener envelopment. Hence, the accurate spatial perception becomes less important for the three successive BRIR parts. Engel *et al.* (2021) already showed that when presenting the direct sound with high spatial resolution, the spatial order of the reverberation part can be reduced significantly. However, the study by Engel *et al.* (2021) is based on the order-limited SMA renderings, and they only examined the direct sound and reverberation separately. In this study, we investigate how the limitation of the spatial resolution of each single BRIR component affects the overall perception of the binaural auralizations.

In two listening experiments, we determined the minimum number of sampling points required to achieve the auralizations indistinguishable from a reference auralization based on the high spatial resolution measurements. The grid order N_{grid} is a suitable parameter to scale the number of sampling points and determine the minimum thresholds. We evaluated each part of the BRIRs separately to investigate if N_{grid} has a varying influence on the different BRIR parts. In two adaptive forced choice ABX listening experiments, the participants had to compare the horizontally head-tracked dynamic binaural synthesis based on the measurements on the sparse grids with orders $N_{\text{grid}} = 1$ to $N_{\text{grid}} = 28$ to a high-spatial resolution reference based on the measurements on a 29th-order grid. With experiment 1, we examined the spatial order of SMA measurements, and with experiment 2, which has partly been published in Lübeck *et al.* (2020b), we examined the spatial order of the dummy head measurements. With these experiments, we tested our hypotheses that (1) the dummy head auralizations require a lower spatial order, i.e., less measured sampling points than the SMA auralizations to be perceptually indistinguishable to the high-resolution reference auralizations and (2) the early reflections and reverberation require a significantly lower spatial order, i.e., less sampling points than the direct sound. Although the study and test design are motivated by our two main hypotheses, we were further interested in how various aspects of the binaural reproduction affect the required spatial resolution. We, hence, performed a broad exploratory listening experiment involving different rooms, audio contents and source positions.

II. THEORY

This section briefly summarizes the two methods for accquiring the binaural signals examined in the listening experiments. First, we will outline the concept of binaural synthesis of the SMA data, which is followed by the SH interpolation of the dummy head measurements. Finally, the signal processing of both methods and the undersampling errors resulting from the limited spatial resolution are compared.

A. Binaural synthesis from SMA captures

The signal processing for calculating binaural signals from the SMA captures has been intensively discussed in the literature, and for more details, the reader is referred to, for expample, Bernschütz (2016) or Rafaely (2015).

The sound field S, which has been sampled on the spherical surface Ω of a microphone array with a radius r, is transformed to the SH domain, applying the spatial Fourier transform (SFT) (Williams, 1999)

https://doi.org/10.1121/10.0009277

$$S_{nm}(\omega) = \int_{\Omega} S(\phi, \theta, \omega)\, Y_n^m(\theta, \phi)^*\, \mathrm{d}A_{\Omega}, \tag{1}$$

where ϕ is the horizontal angle ranging from 0 to 2π, θ is the vertical angle ranging from 0 to π, ω is the angular frequency, and dA_{Ω} is an infinitesimal surface element of Ω. Y_n^m are the surface SH of a certain degree n and mode m, and $(\cdot)^*$ denotes the complex conjugate. With a set of order-dependent radial filters d_n, the radial portion, which is introduced by the SMA body, is removed from the sound field. In this way, the sound field S can be decomposed into a continuum of plane waves D, impinging from all directions, which is known as the plane wave decomposition,

$$D(\phi, \theta, \omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n\, S_{nm}(\omega)\, Y_n^m(\phi, \theta). \tag{2}$$

A HRTF $H(\phi, \theta, \omega)$ is the spatiotemporal transfer function of a plane wave to the listeners' ears. Weighting every sound field plane wave D with the corresponding HRTF from that direction and integrating over the entire surface yields the binaural signals $B(\omega)$, which a listener would be exposed to at the point of the SMA,

$$B(\omega) = \frac{1}{4\pi} \int_{\Omega} H(\phi, \theta, \omega)\, D(\phi, \theta, \omega)\, \mathrm{d}A_{\Omega}. \tag{3}$$

Because the mathematical representation is the same for the left and right ears, for simplification, we omitted the related subscripts throughout this paper. The real-world microphone arrays sample the sound field at discrete positions with a limited number of microphones Q. Consequently, the integrations in Eqs. (1) and (3) become a finite summation, and the plane wave decomposition can solely be calculated up to a certain SH order (Rafaely, 2015). The perceptual consequences of the order-limited plane wave decomposition on the binaural synthesis are mainly degradations of the localization and spaciousness as well as the spectral distortions. These artifacts are discussed, for example, in Ben-Hur *et al.* (2018) or Lübeck *et al.* (2020b) in more detail.

The discretization of Eq. (3) also implies that the sound field can only be decomposed into a limited number of plane waves for the discrete directions. Convolving the limited number of plane waves with the respective head-related impulse responses (HRIRs) is known as the *virtual loudspeaker approach* (Jot *et al.*, 1999). As shown by Bernschütz *et al.* (2014), Ben-Hur *et al.* (2018), or Zaunschirm *et al.* (2018), it is beneficial to perform the convolution with the HRIRs for the plane wave directions on a grid of matched order. The virtual loudspeaker approach can be regarded as the baseline method for binaural decoding, which is why we applied this method in this study.

B. SH interpolation

Nowadays, it is very popular to interpolate the sparsely measured HRTF sets to dense sets in the SH domain (Arend *et al.*, 2021; Aussal *et al.*, 2013; Ben-Hur *et al.*, 2019a; Pörschmann *et al.*, 2020). As this method can be transferred to the BRIRs, we applied the SH interpolation to the sparsely measured dummy head BRIRs in this study. The binaural room transfer functions (BRTFs) are transformed to the spatially continuous SH domain using the SFT [Eq. (1)]. With the inverse spatial Fourier transform (ISFT),

$$B(\phi, \theta, \omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} B_{nm}(\omega) Y_n^m(\theta, \phi), \tag{4}$$

the BRTFs for the arbitrary directions can be calculated. Again, a limited number of sampling points negatively affects the SH representation and introduces audible artifacts. According to Rafaely (2015), these undersampling artifacts increase in magnitude above the spatial aliasing frequency $f = Nc/2\pi r$, where c is the speed of sound.

C. Comparison of undersampling errors

For the SMA captures and BRIR SH interpolation, the grid order limits the SH presentation and mainly degrades the high-frequency components. When interpolating the BRIRs measured with a dummy head, each individual measurement has the maximum spatial resolution and the accurate timbre and, therefore, accurately encodes the entire room information. Here, the artifacts arise from a single back-and-forth SFT, which introduces the SH interpolation errors (Ben-Hur *et al.*, 2018).

On the other hand, the SMA renderings are based on an undersampled sound field, which suffers from spatial aliasing. Moreover, the order-limited SFT of the sound field further impairs the SH representation, leading to plane waves with impaired spectra and blurry spaciousness. This plane wave sound field of limited spatial resolution is then convolved with a set of HRIRs of matched order.

Thus, for both of the methods, the spatial order of the sampling grid degrades the SH presentation of the sound field, resulting in the same artifacts on the rendering side. However, for the SMA renderings, additional spatial aliasing artifacts arise when capturing the sound field. This becomes mathematically clear when considering a sound field consisting of a single plane wave. Substituting a single (ideal) plane wave into Eq. (3) and rearranging it yields a perfectly sampled HRTF from the direction of the plane wave (the derivation can be found in Appendix A). To illustrate this, Fig. 1 depicts the different influences of the spatial undersampling.[1] It shows the binaural signals calculated from a single plane wave impinging from the frontal direction. For the following, we employed the Neumann KU100 HRTF set provided by Bernschütz (2013). As a reference, a HRTF directly measured for the frontal direction is shown as a dark gray curve. As an HRTF describes the transformation of one plane wave to the human ears, this represents the ideal case of an artifact-free rendered plane wave from the corresponding direction. The binaural signal depicted as the red curve represents the real-world SMA case. For this, we simulated the plane wave, which was spatially sampled

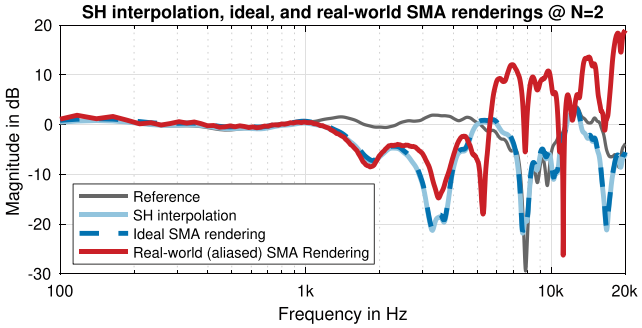


FIG. 1. (Color online) The binaural signals, resulting from the SMA renderings of a simulated plane wave impinging on an ideal virtual SMA (bright blue line), impinging on a real-world SMA with six microphones (red line), and resulting from the second-order SH interpolation of the Neumann KU100 HRTF set (dashed blue line).

at positions of a second-order Lebedev grid. Based on this sampled plane wave, we calculated the binaural signals as described in Sec. II at an order of $N = 2$. The dashed blue curve illustrates the binaural signal resulting from an ideal SMA. For this, again, we simulated an ideal plane wave, which was not sampled by a SMA, and applied the binaural rendering (again, at an order of $N = 2$). These binaural signals are not affected by the spatial aliasing, which occurs when spatially sampling the sound field with the SMA. They are only impaired by the interpolation errors introduced by the order-limited SH processing. Last, the binaural signal depicted as the bright blue line shows the SH interpolation case. For this, we resampled the HRTF set to a second-order Lebedev grid. We then transformed the resampled HRTF set to the SH domain at an order of $N = 2$ and inverse transformed it for the frontal direction. It can be seen that the binaural signal resulting from the ideal SMA (dashed blue curve) is identical to the signal from the SH interpolation (bright blue curve). The real-world SMA binaural signal is notably more impaired. This example shows that on the rendering side, the SH interpolation and SMA rendering are impaired by the order-limited SH processing to the same extent. The binaural signals from the real-world SMA renderings, however, are further impaired by the spatial aliasing on the capturing side. The aliasing and truncation errors are mathematically derived in Ben-Hur *et al.* (2019b). It is worthwhile to mention that for the ideal and real-world SMA renderings in this example, we used ideal radial filters. Thus, the plot only shows the nonideal behaviour of the real-world SMA renderings in terms of the undersampling errors and neglects the constraints of the nonideal radial filters such as the soft-limited radial filters (Bernschütz *et al.*, 2011b).

III. EXPERIMENT 1: AURALIZATION OF SMA DATA

In the first listening experiment, we determined the minimum grid order of the sparse SMA sampling grids, which results in binaural auralizations that are indistinguishable from the auralizations of the SMA data measured on a 29th-order grid. We determined this minimum order as the point of subjective equality (PSE) in an adaptive ABX listening experiment.

A. Method

1. Participants

A total of 36 participants, 29 males and 7 females with a mean age of 24.6 years old [standard deviation (SD), 5.4 years], took part in the listening experiment. Most of them were media technology students, and all of them had self-reported normal hearing.

2. Setup

We applied the dynamic binaural synthesis using the SoundScape Renderer (Geier *et al.*, 2008, 2019). It convolves a set of BRIRs with arbitrary anechoic input signals according to the listener's head orientation, which was tracked with a Polhemus Fastrak (U - 05446-Vermont, US) at a sampling rate of 120 Hz. The experiments were performed in the anechoic chamber of TH Köln with a background noise level of less than 20 dB(A). We used an RME Fireface UFX (D-85778 Haimhausen, Germany) as a digital-analog converter at 48 kHz and a buffer size of 256 samples and Sennheiser HD600 headphones (DE - 30900 Wedemark, Germany) for playback with a playback level of about 66 dB(A). We equalized the binaural chain of the Neumann KU100 dummy head (DE-10117 Berlin, Germany) and Sennheiser HD600 headphones using a 2048 tap minimum phase compensation filter designed according to a regularization method proposed in Erbes *et al.* (2017). The test was implemented and performed with the MATLAB software Scale (Giner, 2013).

3. Stimuli

a. Employed data. For the listening experiments in this study, we used the SMA impulse responses captured in four different rooms at the WDR broadcast studios (Stade *et al.*, 2012). The impulse responses were sampled on a 1202 node Lebedev grid, which allows the SH representation up to the 29th order. At an order of $N = 29$, the spatial aliasing and SH order truncation artifacts can be neglected up to approximately 18 kHz (with a radius of 0.0875 m and a speed of sound of 343 ms^{-1}; Rafaely, 2015, p. 80). Thus, the 29th-order SMA captures are well-suited as the high-spatial resolution ground truth in this study. The database consists of the measurements in the four different rooms with different reverberation times as presented in Table I. For synthesizing the binaural signals, again, we used the Neumann KU100 HRTFs (Bernschütz, 2013).

Because we intended to investigate the spatial resolution separately for the three parts of the BRIR, i.e., the direct sound, early reflections, and reverberation, we defined the transition times of these parts as follows. The direct sound corresponds to the duration of a HRTF measured in anechoic conditions (Blauert, 1996; Møller *et al.*, 1995; Zahorik, 2002) and is approximately 2.5–3.5 ms. For some cases, it is difficult to separate the first floor reflection from

JASA
https://doi.org/10.1121/10.0009277

TABLE I. The RT_{60} and the transmission times at which the early reflections and the reverberation part start for all of the rooms examined. RT_{60}, The reverberation time 60 (500 Hz and 1 kHz).

Room	RT_{60}	Early reflections starting time	Reverberation starting (mixing) time
Control room 1 (CR1)	<0.25 s	3.5 ms after onset	71.02 ms
Control room 7 (CR7)	<0.25 s	3.5 ms after onset	39.03 ms
Small broadcast studio (SBS)	0.9 s	3.5 ms after onset	43.34 ms
Large broadcast studio (LBS)	108 s	3.5 ms after onset	46.22 ms

the direct sound, which is why we decided to define the duration of the direct sound for all rooms as 3.5 ms. The direct sound is followed by a number of early reflections, and at the so-called mixing time, the number of reflections has increased such that the sound pressure is equally distributed over the entire room, and the sound field can be considered as diffuse. Different methods to estimate the mixing time have been proposed in the literature. A comprehensive comparison and evaluation of the various methods has been presented in Lindau *et al.* (2010). In this study, the mixing times have been estimated with a procedure introduced by Abel and Huang (2006) and as proposed in Lindau *et al.* (2010), which is averaged across the left and right ear signals for the frontal direction with a window length of 20 ms and a safety margin of 100 samples. The resulting mixing times are presented in Table I.

b. Simulation of sparse measurements. The reference BRIRs were calculated from the 29th-order Lebedev grid SMA impulse responses according to Eq. (3). To simulate the SMA measurements on the sparse grids, we spatially resampled the 29th-order grid to order $N = 1$–28 Gauss grids by interpolation in the SH domain. The SMA impulse responses are transformed to the SH domain up to the maximum order of 29. Subsequently, the ISFT [Eq. (4)] is applied to yield the 28 SMA impulse response sets defined for the sampling directions according to $N = 1$–28 Gauss grids. This procedure results in sparse SMA impulse responses, which suffer from spatial aliasing and SH order truncation, as would be the case with measurements with the real-world SMAs. In contrast, truncating the SH order series of the SH representation of the 29th-order grid would solely lead to the SH order truncation artifacts. Gauss grids are rather inefficient, i.e., they need a relatively large number of sampling points to resolve certain SH orders. However, in contrast to the more efficient grids, such as the Lebedev or Fliege grid, Gauss grids are defined for every grid order. Therefore, we decided to use Gauss grids to scale the spatial resolution in steps of one order. The influence of different grids has been discussed, for example, in Rafaely (2015, Chap. 3), Zotter (2009, Chap. 4.2), or Bernschütz (2016, Chap. 3.2.2).

From each of the resampled sparse SMA impulse responses, we calculated the BRIRs as described in Sec. II up to the corresponding SH order. According to the virtual loudspeaker approach, for each BRIR rendering, the plane wave components were calculated for the Gauss grids of corresponding order and weighted with the HRTFs for these directions. The plane wave components were calculated with radial filters with a 20 dB soft limit (Bernschütz *et al.*, 2011b). As an alternative to the rotation of the entire sound field in the SH domain, the BRIRS for the different listener head orientations can be synthesized by accounting for the head orientation when selecting the HRTFs for the plane wave incident directions [in Eq. (3)]. This procedure has been discussed by Bernschütz (2016, p. 66) in more detail. Besides the reference BRIRs, we calculated 28 BRIR sets based on 28 SMA impulse response sets of order 1–28. Each of the BRIR sets were calculated for 360 directions in 1° steps along the horizontal plane. However, in the listening experiment, the binaural synthesis was adapted according to the listeners' head orientations only for ±60° along the horizontal plane to save the working memory. The signal processing was performed in MATLAB using the SOFiA toolbox (Bernschütz *et al.*, 2011a). The block diagram in Fig. 2 presents an overview of the signal processing.

c. Splitting the BRIRs. To determine the minimum grid order for each BRIR component separately, we split each BRIR set at the transmission times specified in Table I. This resulted in the direct sound part, consisting of the first 3.5 ms, the early reflection part up to the mixing time, and the final reverberation part. Subsequently, we recomposed the sparse BRIR sets with a reduced spatial resolution in (a) just the reverberation (REV) part, (b) just the early reflections and reverberation parts (ER), and (c) in all three parts, resulting in the BRIRs being completely reduced in their spatial resolution (DS). To ensure the artifact-free recomposition, we applied linear fading over 128 samples between the parts. In the following, these BRIRs are denoted as

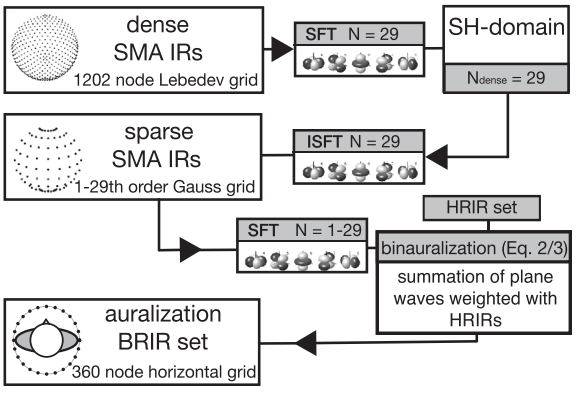


FIG. 2. The block diagram of the signal processing for the generation of the BRIRs based on the sparse SMA impulse responses examined in experiment 1.

Rev-BRIRs (reduced spatial resolution starting at the mixing time), *ER-BRIRs* (reduced spatial resolution starting at the early reflections), and *DS-BRIRS* (reduced spatial resolution starting at the direct sound).

d. Test signals and sound source positions. Because several studies (Ahrens and Andersson, 2019; Arend *et al.*, 2021; Pörschmann *et al.*, 2019; Zaunschirm *et al.*, 2018) showed that lateral sound sources are perceptually more critical than frontal sources, we presented one sound source from the side at $\phi = 270°$. Further, we examined a second position at $\phi = 30°$ to present a frontal source that induces the interaural time and level differences. As anechoic test signals, we used a pink noise burst with a length of 0.75 s (including 10 ms cosine-squared onset/offset ramps) and a male speech sample, consisting of a short German sentence with a length of 1.5 s.

4. Procedure

The ABX three-interval/two-alternative forced choice (3I/2AFC) test design is a simple, robust, and widely used paradigm in psychophysics. In combination with the adaptive one-up one-down staircase procedure (Kingdom and Prins, 2010; Levitt, 1971, Chap. 3 and 5), it is well-suited to determine the so-called PSE (Meese, 1995). The PSE, also denoted as the threshold of recognition, is the 50% point on the psychometric function. It defines the point at which no relevant differences can be detected anymore, as in the present case, the differences between the BRIR auralizations based on the sparse and dense SMA measurements.

According to the ABX test paradigm, three intervals (A, X, and B) were presented to the participants in each trial. Two of the intervals consisted of the same stimuli (i.e., auralizations based on the BRIR with the same spatial resolution). Either the reference stimulus (based on the high-resolution BRIR) or the stimulus of the lower-resolution BRIR was assigned randomly to the middle interval X. Accordingly, either A or B consisted of the same stimuli as X. This assignment ensures the (1) direct comparison of the stimuli with different spatial resolutions and (2) that either the lower or higher resolution was presented two times randomly. After the presentation of the three intervals A, X, and B, the participants were asked to decide if A or B equaled X by pressing the corresponding button on the experiment graphical user interface (GUI).

Following the adaptive one-up-one-down staircase method, if the participants were correct and could indicate the difference between the intervals, the stimulus based on the BRIR with the next higher spatial order was assigned in the next trial. If they gave a wrong answer, i.e., they could not indicate any differences, the BRIRs based on the next lower grid were picked for the next trial. Each run started with the low resolution stimuli of order $N_{\mathrm{grid}} = 1$ and was terminated after 12 reversals. One reversal is defined as a correct decision followed by a wrong decision or vice versa. Each A, B, X sequence was automatically played back one time and, during the playback, the participants were free to move their heads if it helped them to distinguish between the sparse and dense BRIR auralizations.

According to a $4 \times 3 \times 2 \times 2$ mixed factorial design with the between-subject factor room (CR1, CR7, SBS, LBS) and the within-subject-factors BRIR component (DS, ER, REV), source position ($\phi = 30°$, $\phi = 270°$), and test signal (noise, speech), the participants were divided into four groups with nine participants each. The participants from each group performed 12 runs in total (3 BRIR components, 2 source positions, and 2 test signals). To ensure that the subjects fully understood the task, each participant was introduced to the experimental design at the beginning of the experiment. Afterward, each participant had to perform two training runs to get familiar with the test procedure. The training runs consisted of the speech test signal at position $\phi = 30°$ and the noise test signal at position $\phi = 270°$ for the corresponding room. The training runs were completed after four reversals or five wrong decisions at the lowest spatial resolution. The duration of the experiment varied depending on the group and participants. On average, the experiment took about 50 min, including a short break.

B. Data analysis

The PSEs were determined as the averaged grid order N_{grid} over the last nine reversals. We, thus, omitted the first three reversals. Because the rendering requires an integer number for the grid order N, we employed discrete values for the statistical analysis. According to Yap and Sim (2011) or Bee Wah and Mohd Razali (2011), the Shapiro-Wilk test is powerful for testing the assumption of the normality distribution, which is why we decided to use it in this study. Moreover, Chen *et al.* (2017) presented a comprehensive comparison of the different adjustments for multiple testing. We decided to use the rather conservative Bonferroni method for all of the adjustments. A Shapiro-Wilk test with Bonferroni correction showed no violations of the assumption of the normality distribution of the data. Therefore, we analyzed the PSEs with a four-way mixed analysis of variance (ANOVA) with the between-subject factor room (CR1, CR7, SBS, and LBS) and within-subject factors BRIR component (DS, ER, REV), test signal (noise, speech), and source position (30°, 270°). A Mauchly test for the sphericity revealed that for the factor BRIR, the component sphericity was not met, which is why we applied the Greenhouse-Geisser correction where applicable. Girden (1992) proposed to use the Greenhouse-Geisser correction for $\varepsilon \leq 0.75$. The ANOVA for experiment 2 revealed $\varepsilon \leq 0.75$. Therefore, we decided to apply the Greenhouse-Geisser correction for all of the applied tests. For a more detailed analysis, we further applied the various *post hoc* Bonferroni corrected independent-samples t-tests.

C. Results

A graphical overview of the results is presented as boxplots in Fig. 3. First, it can be seen that the PSEs become

https://doi.org/10.1121/10.0009277

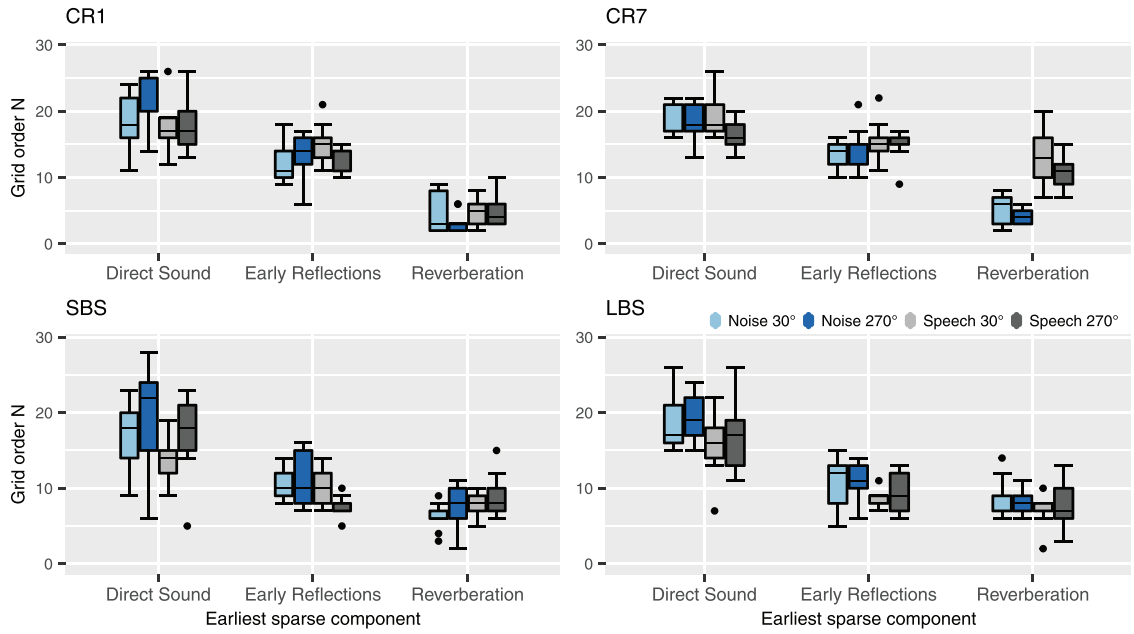


FIG. 3. (Color online) The interindividual variation in the determined PSEs (grid orders N) for the tested rooms CR1, CR7, SBS, and LBS with respect to the earliest sparse component (x axis) are shown for experiment 1. The sound source position and test signal are depicted separately as indicated by the colors. Each box shows the interquartile range (IQR), median value (black line), outliers (black points), and black whiskers, displaying the 1.5 × IQR below the 25th or above the 75th percentile. Note that in some cases, the median is exactly on the upper or lower IQR. The median of the CR1 direct sound noise at 270° is 20, the median of theCR1 early reflections speech at 270° is 17, the median of the CR7 early reflections noise at 270° is 12, and the median of the LBS reverberation noise at 30° is 9.

smaller for the successive BRIR components. The reverberation part of the CR7 room is an exception as it shows the relatively large PSEs for the speech test signal. Furthermore, it can be observed that for the direct sound part, the medians of the PSEs for the more critical test signal noise are always higher. For the more reverberant rooms SBS and LBS, the medians of the PSEs for the lateral sound source position at 270° are higher than those for the frontal 30° condition. The boxplots further indicate the PSE outliers. The highest PSE of 28 was detected for the direct sound part of the SBS at the lateral position and noise test signal. It is worth mentioning that this PSE of 28 is below the upper whisker and was, thus, not indicated as an outlier.

The results of the four-way mixed ANOVA are shown Table II of Appendix A. The significant main effect of the BRIR component together with the observation from Fig. 3 indicates a strong dependency of the BRIR component on the required grid order. The ANOVA further revealed a significant main effect of the room as well as the interaction effects of room × BRIR component, room × source position, and room × signal. These significant differences of the room might be due to the exception of CR7 in the reverberation part. The interaction of BRIR component × signal shows that the test signal has varying influence on the PSE for the different BRIR components. Although the position is not a significant main effect, it has a varying influence with respect to the room as indicated by the interaction effect of position × room. Last, we found the significant interaction effect of room × BRIR component × signal. A *post hoc* power-analysis with G*power (Erdfelder *et al.*, 2009) based on the calculated Cohens' f values revealed an achieved power $\geq$0.9 for all of the significant effects.

To further investigate the significant effect of the room, we applied a series of *post hoc* independent-samples t-tests between the pooled data of each room. Only the t-tests between the data for the room CR7 and SBS and CR7 and LBS were significant, which supports the assumption that the effect of the room is due to the exception in CR7 (the results of all of the t-tests are displayed in Appendix B 1).

For further inspection of the significant effect of the BRIR component, Fig. 4 displays the mean values for each room and the BRIR component separately pooled over the signal and position. It can be seen that the means between all of the BRIR components vary for each room. Only for the more reverberant rooms SBS and LBS, the visual inspection does not indicate a clear difference between the early

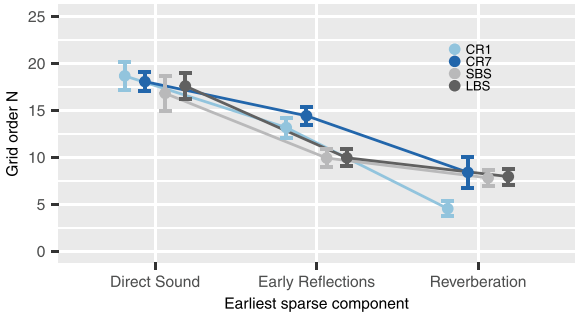


FIG. 4. (Color online) The marginal mean plot with respect to the room and BRIR pooled over the position and signal, including 95% within-subject confidence intervals, is shown for experiment 1.

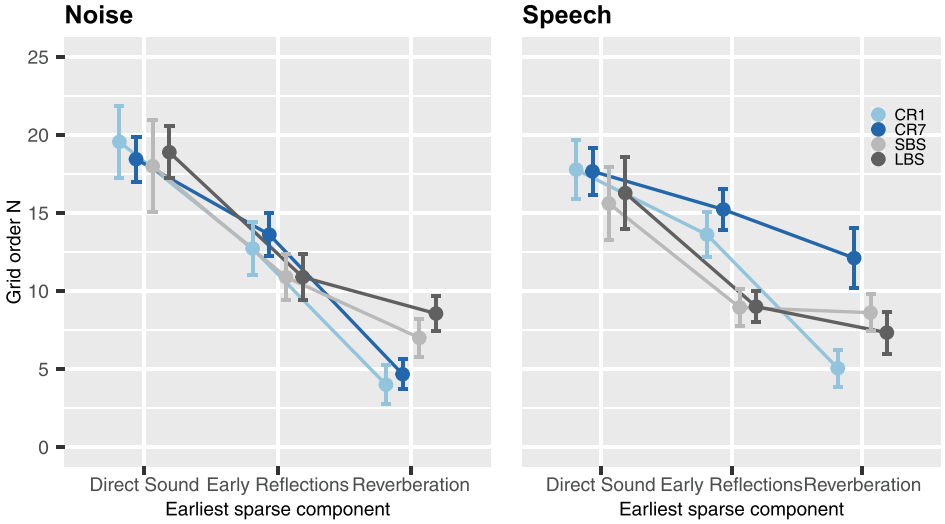


FIG. 5. (Color online) The average PSEs pooled over the signal with respect to the earliest sparse BRIR component (x axis) and room (color) for the frontal and lateral positions separately and the 95% within-subject confidence intervals are shown for experiment 1.

reflections and reverberation part. However, a t-test showed that there is a significant difference between them for both rooms. This suggests that there is a significant difference between each BRIR component for all of the rooms, and the required spatial orders descend over the successive BRIR components. The crossing of the interaction line of the room CR1 with the LBS and SBS illustrates the interaction between the BRIR component and the room. For the later BRIR components, CR1 seem to require significantly less spatial orders than CR7, SBS, and LBS. Probably, this is simply due to the higher estimated mixing time in CR1 (see Table I) and, thus, a shorter reverberation time part. A *post hoc* nested ANOVA involving the PSEs of the direct sound part with the between-subject factor room and the within-subject factors source position and test signal only showed a significant effect of the test signal [$F(1,32) = 8.36$, $p < 0.007$, $\eta_p^2 = 0.21$, $\varepsilon = 1.0$], which suggests that the dependency of the factor room might be attributed to the later BRIR parts.

For further investigation of the interactions of the room × BRIR component, room × signal, and room × position, we performed *post hoc* nested ANOVAs with the within-subject factors BRIR component, source position, and test signal for each room separately (see Tables III–VI in Appendix B 1). Each ANOVA revealed a significant main effect BRIR component. Only for CR7, we found a significant effect of the position. This indicates that CR7 causes the interaction of the room × position. For all of the other rooms, the position has no significant influence on the required spatial order. This can also be seen in Fig. 5, which displays the mean values with respect to the room and BRIR pooled over the signal for both positions separately. For CR7 and LBS, we further found a significant effect of the signal. This indicates that CR7 and LBS cause the significant interaction effect of the room × signal. This is strongly supported by Fig. 6, which displays the mean values with respect to the room and BRIR component pooled over the positions for both signals separately. The reverberation part

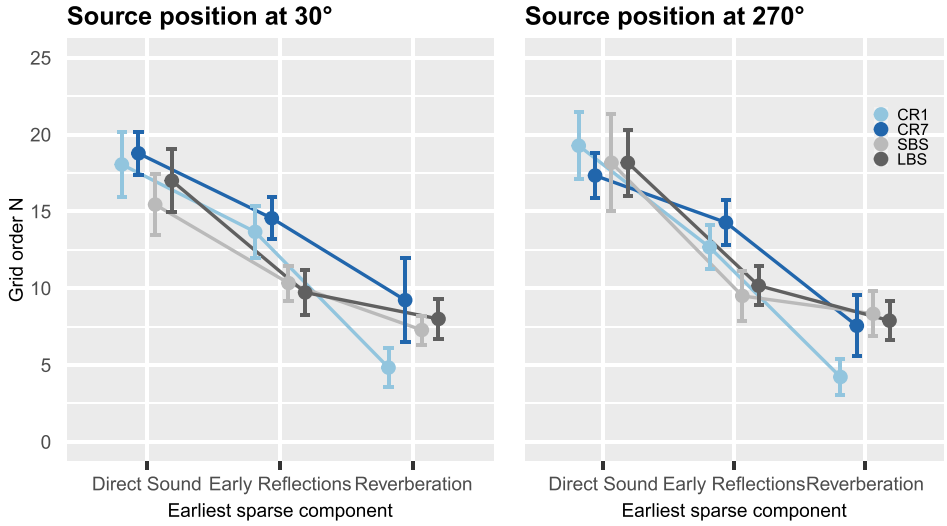


FIG. 6. (Color online) The average PSEs pooled over the position with respect to the earliest sparse BRIR component (x axis) and room (color) for the frontal and lateral positions separately and the 95% within-subject confidence intervals are shown for experiment 1.

https://doi.org/10.1121/10.0009277

of CR7 requires significantly higher orders for the speech than for the noise signal. For CR7, SBS, and CR1, we found the significant interaction BRIR comp × signal. They cause the three-way interaction room × BRIR component × signal. The p*ost hoc* paired *t*-test between both signals for each room and BRIR component separately revealed that for only the reverb part of CR7, the signal leads to a significant difference in the PSEs.

For a better overview and comparison with the results of experiment 2, Table VIII (Appendix B 1) displays all of the mean values across the subjects with respect to the BRIR component, test signal, and source position. Although the factor room had a significant effect, we decided to pool the data of all of the rooms. Thus, the interpretation of the mean values, at least for the early reflections and the reverberation, should be performed with reservation.

Based on the results of experiment 1, we, thus, conclude as follows. The minimum required spatial resolution varies and mostly decreases for the three successive BRIR components, which supports our second main hypothesis. For the direct sound, the average PSEs range from 17 to 20 for the early reflections from 12 to 13 and for the reverberation from 7 to 9. The PSEs of each BRIR component vary significantly. We could observe a significant influence of the auralized room, whereby there are indications that this dependency is evoked by the later BRIR parts of the early reflections and reverberation. At this point, it should be noted that different participants with different experiences, for example, in spatial audio, might result in different test results. Thus, especially for the between-subject test designs, a significant effect of the between-subject factor (such as, in this case, the room) could be due to the participant group.

IV. EXPERIMENT 2: AURALIZATION OF DUMMY HEAD DATA

In the second listening experiment, we investigated the dummy head auralizations. We determined the minimum number of sampling points, which after interpolation in the SH domain results in auralizations that are indistinguishable to the auralizations of the reference measurements on a dense 29th-order grid. The results of this listening experiment have partly been presented in Lübeck *et al.* (2020b). The setup, test design, and procedure were exactly the same as for those in experiment 1.

A. Method

We were interested in a later comparison of the results of both experiments. To have a balanced number of observations, we extended the data by four participants compared to the results presented in Lübeck *et al.* (2020b). A total of 36 participants, 25 male and 11 female, took part in the listening experiment (mean age = 28.1 years old, SD = 7.4 yr). Most of them were media technology students and all had self-reported normal hearing.

The stimuli for this experiment are based on the same impulse responses as in experiment 1. Because the employed database does not contain a full-spherical dummy head BRIR measurement set on a 29th-order grid, the high-spatial resolution references were also calculated from the 29th-order SMA impulse responses for 1202 head orientations of a 29th-order Lebedev grid, according to Eq. (3). The so-computed BRIRs are nearly perceptually equivalent to the BRIRs directly measured with a dummy head as has been extensively discussed and evaluated, e.g., in Ahrens and Andersson (2019) and Bernschütz (2016). Therefore, we considered these BRIR sets as the high-resolution ground truth.

To simulate the sparse dummy head measurements, we transformed these 29th-order Lebedev BRIR sets to the SH domain at the maximum order of 29. Subsequently, applying the ISFT, we resampled the dense set to 28 BRIR sets defined for the sampling directions according to $N = 1$–28 Gauss grids. Finally, for the continuous dynamic binaural synthesis, all of the sparse BRIR sets were transformed to the SH domain again, this time with the corresponding order of the sparse BRIR set, and then resampled to a 360 sampling point grid with 1° steps. We split the BRIRs in the direct sound part, early reflections part, and reverberation part, and recombined them exactly as was done for experiment 1. Again, the binaural synthesis was adapted according to the listeners' head orientations only for ±60° along the horizontal plane. The entire signal processing workflow is illustrated as a block diagram in Fig. 7. In all other aspects, the setup, materials, procedure, and analysis were identical to those of experiment 1.

B. Results

A Bonferroni corrected Shapiro-Wilk test rejected the hypothesis of normal distribution in 1 of 48 conditions. However, the parametric tests are robust to slight violations of normality assumptions (Bortz and Schuster, 2010; Pearson, 1931). Therefore, for the statistical analysis, again, we applied a four-way mixed ANOVA with the between-subject factor room and the within-subject factors BRIR

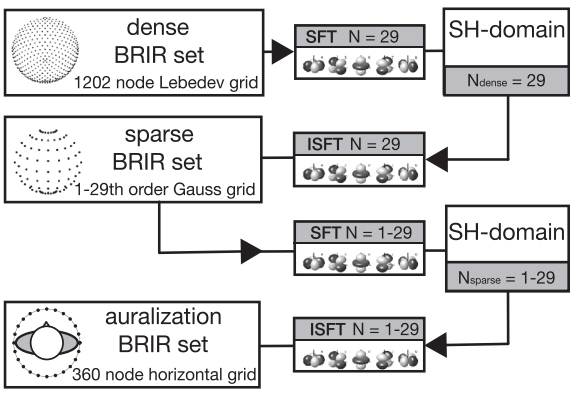


FIG. 7. The block diagram of the signal processing for the generation of the sparse BRIR sets in experiment 2.

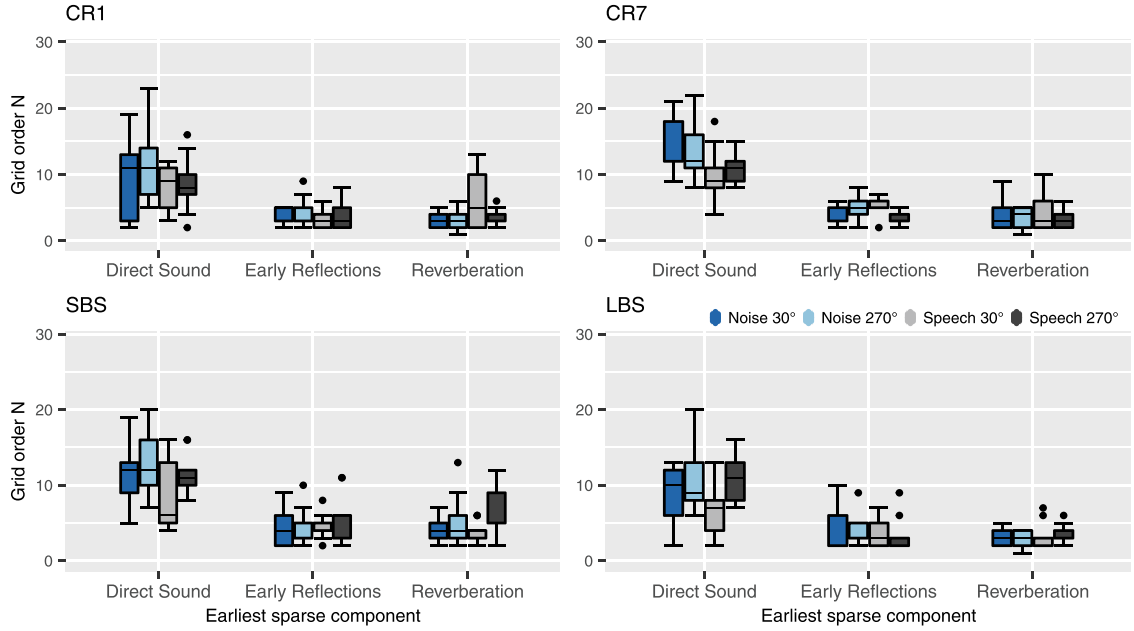


FIG. 8. (Color online) For experiment 2, the interindividual variation in the determined PSEs (SH orders N) for the tested rooms CR1, CR7, SBS, and LBS with respect to the earliest sparse component (x axis). The sound source position and test signal are depicted separately as indicated by the colors. Each box specifies the interquartile range (IQR), median value (black line), outliers (black points), and black whiskers, displaying the 1.5 × IQR below the 25th or above the 75th percentile. Again, in some cases, the median is exactly on the upper or lower IQR. The median of the CR7 direct sound noise at 30° is 12, the median of the CR7 early reflections noise at 270° is 3, the median of the SBS early reflections noise at 270° is 5, the median of the SBS early reflections speech at 270° is 3, the median of the SBS reverberation speech at 270° is 5, and the median of the LBS early reflections noise at 30° is 2, and the median of the LBS early reflections noise at 270° is 3.

component, test signal, and sound source position. Just as with experiment 1, the Mauchly test rejected the assumption of sphericity for the factor BRIR component, and we applied the Greenhouse-Geisser correction. Figure 8 presents an overview of the results of experiment 2. For each room, the PSEs significantly decrease for the BRIRs with a limited resolution of the early reflections and reverberation. Between the early reflections and reverberation, we cannot observe a difference by visual inspection. For all of the rooms, noise was the more critical test signal for the direct sound. This dependency of the test signal seems to become smaller for the early reflections and reverberation part. It can be seen that for all of the rooms, none of the participants could detect differences of the grids with orders higher than 23. The absolute maximum value of all of the rooms was 23 for CR1 with the noise test signal at 270°.

The results of the four-way mixed ANOVA are displayed in Appendix B 2, Table IX. Because we could neither observe any significant main effect of the room nor any interaction effect involving the factor room, we pooled the data over the room for Fig. 9, which supports the results of the ANOVA.

The ANOVA revealed a significant main effect of the BRIR component, which is strongly supported by the observations from Figs. 8 and 9. The significant effect of the source position also matches the observation from Figs. 8 and 9 and shows that the lateral source positions mostly required higher grid orders than frontal grid orders. Furthermore, we found significant interaction effects of the BRIR component × source position and BRIR component × signal. This indicates that the signal and source position have varying influences on the PSE with respect to the BRIR component, which was already observed in Figs. 8 and 9.

To disentangle the interaction effects, we applied a series of Bonferroni corrected independent-samples t-tests between the data of positions 1 and 2 for each BRIR component separately. For the direct sound only, we found a significant difference between the frontal and lateral source positions. We performed the same t-tests between the noise and speech test signal and also found that only for the direct sound, the PSEs of the noise and speech signal differ significantly. This supports the assumption that for the direct sound, the source position and signal have an influence on the PSE but this is not the case for the later BRIR components.

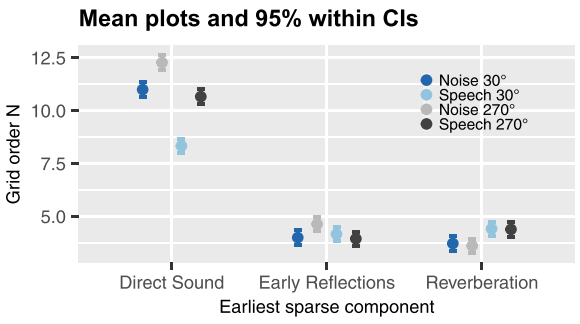


FIG. 9. (Color online) For experiment 2, the average PSEs pooled over all of the rooms with respect to the earliest sparse BRIR component (x axis) are shown. The 95% within-subject confidence intervals were calculated according to Jarmasz and Hollands (2009) and Loftus (1994). The test signal and sound source positions are displayed separately as indicated by the colors.

Moreover, we performed the pairwise t-tests between all of the BRIR components and found that the PSEs of the early reflections and reverberation parts are not significantly different. The PSEs of the direct sound significantly differ from both.

Similar to experiment 1, Table X in Appendix B 2 shows the mean values across the subjects with respect to the BRIR component, test signal, and source position.

The results of experiment 2 lead to the following assumptions. As similar to experiment 1, the required grid order decreases significantly for the successive BRIR parts but with notable smaller PSEs: 9–13 for the direct sound and 4–5 for the early reflections and the reverberation. Together with the results of experiment 1, this proves the second main hypothesis. We could not detect a significant difference for the PSEs in the early reflections and reverberation. For the direct sound part, the test signal and source position have a statistical influence; for the early reflection and reverberation part, this influence vanishes. We did not observe any statistical significance of the room. Because in experiment 2, the ANOVA did not reveal any significant three-way interaction, we do not show the estimated marginal mean plots as were shown for experiment 1.

The results plots in Figs. 3 and 8 as well as the averaged PSEs in Tables VIII and X suggest that the PSEs for experiment 1 are significantly higher than those for experiment 2. To prove that these differences are statistically significant, we further applied a series of independent-samples t-tests (two-sided) with the Bonferroni correction. The t-test comparing the pooled PSEs from experiment 1 and experiment 2, respectively, revealed a significant difference in the estimated PSEs between both of the experiments and, thus, between the minimum required SH order for the SMA and dummy head BRIRs [$t(862) = 17.175$, $p < 0.001$, $d = 1.17$]. The separate pairwise comparisons of the PSEs estimated in experiments 1 and 2 for the three different BRIR components (DS, ER, and REV) showed significant differences for all of the BRIR components [DS, $t(286) = 13.79$, $p < 0.001$, $d = 1.63$; ER, $t(286) = 22.661$, $p < 0.001$, $d = 2.67$; REV, $t(286) = 8.78$, $p < 0.001$, $d = 1.03$]. This suggests that the first main hypothesis is also valid.

V. GENERAL DISCUSSION

A. Comparison of both experiments

Both of the experiments show that the required grid order varies and mostly decreases for the later BRIR parts. As expected, the dummy head SH interpolation requires significantly lower grid orders than the SMA renderings. Moreover, it is noticeable that in experiment 1, there is a significant difference between the PSEs of the early reflections and reverberation, whereas for experiment 2 this is not the case. Further, the results of experiment 2 show that for the direct sound part, the test signal and source position have a significant effect, whereas this influence vanishes for the later BRIR part. We conclude that the SH interpolation of the dummy head data mainly affects the direct sound part and seems to produce fewer perceptual artifacts in the later BRIR components. On the contrary, limiting the order of the SMA renderings seems to have more impact on the synthesis of the later BRIR parts. Certainly, it is due to the different signal processing applied for both methods. One explanation could be that the SMA renderings synthesize the BRIRs with a superposition of (order-limited) plane waves. The limitation of the plane wave decomposition results not just in impaired plane waves but also in fewer directions of the impinging plane waves, which might introduce the comb-filtering artifacts. We, thus, assume that in contrast to the dummy head SH interpolation, the SMA renderings require relatively high SH orders to synthesize the diffuseness of the sound field and its timbre. This could also be an explanation for the room dependency of the SMA renderings, which was not observed for the dummy head SH interpolation.

B. Comparison to previous studies

Ahrens and Andersson (2019) presented a listening experiment, comparing the dummy head auralizations to the order-limited SMA auralizations. They found that mostly above orders of eight, the perceptual differences decrease. Our experiments show that even up to an order of 28, the perceptual differences persist. However, in Ahrens and Andersson (2019), the participants rated the difference in terms of the spaciousness and timbre on a quality scale, whereas we examined the overall indistinguishability in experiment 1.

Recently, Engel *et al.* (2021) and Engel *et al.* (2019) presented a comprehensive investigation on the different Ambisonic-based binaural renderers, in which the direct sound and the reverberation were rendered separately. They found that when the direct sound is auralized with the high spatial resolution, the listeners could hardly distinguish between the first-, second-, third-, and fourth-order binaural reverberations. However, the participants compared to a reference rendering at a SH order of four. Our experiment 1 shows that when comparing to a high-order reference, in our case $N_{\text{grid}} = 29$, orders 6–9 are necessary for the indistinguishability of the reverberation. Unlike Engel *et al.* (2021) and Engel *et al.* (2019), who truncated the SH order of the Ambisonics representation, we resampled the SMA data to the sparse grids. However, the general findings of both of the studies are similar: The reverberation part in the binaural auralizations can be rendered with lower SH orders than the earlier parts. In addition, Engel *et al.* (2021) and Engel *et al.* (2019) did not distinguish between the required spatial orders of the dummy head and SMA data. Our study shows that these orders are significantly different.

In the past studies, there are inconsistent and even contrary observations regarding the room dependency of the SMA renderings. Bernschütz (2016, p. 224), as well as Ahrens and Andersson (2019), did not detect any statistical differences across the rooms. On the contrary, Ahrens *et al.* (2017) reported that reverberant rooms require higher SH

orders than less reverberant rooms for the indistinguishability compared to the dummy head auralizations. Engel *et al.* (2021) also found that certain room characteristics affect the required spatial order. In line with this finding, the ANOVA for our experiment 1 involving all of the rooms showed a significant effect of the room. However, the ANOVA that only considers the direct sound condition did not reveal the room as a significant effect. Furthermore, there are indications that the significant effect of the room is only evoked by the later BRIR components of the room CR7. For the reverberation part of CR7, we found relatively high PSEs. Interestingly, they were detected for the speech signal in the front, although it is the less critical signal at the less critical position. Visual inspection of the impulse responses around this part did not yield any anomalies such as strong reflections. It is worth mentioning that for experiment 2 in the reverberation part of the CR1, which has a comparable RT_{60}, we could observe a similar tendency, again, for the speech signal in the front. However, this observation was not indicated as significant by the ANOVA. We could not find a clear explanation for this interaction of the dry rooms and the speech signal in the reverberation part.

We assume that, in general, there is a weak influence of the room on the binaural SMA renderings, which is certainly more significant in the later BRIR parts.

Experiment 2 indicates that the SH interpolation of the dummy head data is not dependent on the room. In this context, it is interesting to compare experiment 2 to the studies of Pike (2019, Chap. A.8] or Arend *et al.* (2021). Pike (2019) compared the auralizations of the anechoic HRTFs, interpolated in the SH domain to the HRTFs directly measured at that direction. They showed that above an order of 35, no differences were audible anymore. Arend *et al.* (2021) conducted a similar experiment as the present experiment 2 but just for the anechoic HRTFs and reported the PSEs between 13 and 25 for the frontal and lateral noise and speech sound sources. In contrast, our study revealed the PSEs between 9 and 13 for the direct sound. It can, thus, be assumed that in the presence of early reflections and reverberation, the artifacts in the direct sound are perceptually less relevant. Therefore, the SH interpolation of the dummy head impulse responses, which also encode the reflections and reverberation, requires significantly smaller SH orders than the SH interpolation of the dummy head impulse responses, which encode only the direct sound in the anechoic conditions, i.e., the HRIRs.

To determine the PSE thresholds for both of the methods, we used a baseline approach, i.e., the virtual loudspeaker method to synthesize the BRIRs from the SMA data, and the classical SH transform for the interpolation of the BRIRs. However, in the last years, several approaches have been developed to perceptually improve the binaural renderings of the SMA captures, for example, as discussed in Zaunschirm *et al.* (2018) or Lübeck *et al.* (2020a). Also, the interpolation of the BRIRs in the SH domain could be improved by the spectral equalization, matrix regularization, or time-alignment approaches. Therefore, it should be noted that different decoding or interpolation methods may lead to different thresholds. However, the baseline methods allow us to keep the signal processing for both of the methods similar (see Sec. II) and determine the generally valid but rather conservative thresholds.

VI. CONCLUSION

In this paper, we presented two listening experiments with the aim of finding the minimum required spatial orders of the SMA and dummy head BRIR measurements, which result in an auralization that is indistinguishable from a high-resolution reference. We applied the dynamic binaural synthesis, which was adapted only with respect to the horizontal head orientation of the listener. The found thresholds may shift for the full-spherical auralizations. For the horizontally head-tracked auralizations, we could show that the BRIR components encoding the early reflections or the reverberation for the dummy head data and SMA data require fewer sampling points than the direct sound component. Furthermore, the dummy head impulse responses require lower orders than the SMA impulse responses to achieve the perceptually similar binaural auralization. Last, the room has no influence on the interpolation of the dummy head BRIRs in the SH domain, whereas for the SMA renderings, it has an influence. The thresholds can be used to further simplify the data acquisition of the binaural rendering. Furthermore, the computational effort can be reduced enormously when rendering the direct sound, early reflection, and reverberation separately. In this study, we determined the thresholds in terms of the indistinguishability. It can be assumed that the quality-based listening experiments would lead to significantly lower spatial orders.

ACKNOWLEDGMENTS

The authors would like to thank all of the participants for their support as well as Melissa Ramírez and Kai Altwicker for assistance in conducting the listening experiments. We are grateful to the two anonymous reviewers for their constructive comments on previous versions of this manuscript. The work was funded by ERDF (European Regional Development Fund) under the funding reference code EFRE-0801444.

APPENDIX A: UNDERSAMPLING ERRORS IN DUMMY HEAD AND SMA RENDERINGS: MATHEMATICAL DERIVATIONS

In the following, it is mathematically shown that if undersampling artifacts due to the discrete sampling of the sound field are neglected, the spatial interpolation of the dummy head data in the SH domain is equivalent to the binaural rendering of the SMA data. The interpolation of a HRTF set $H(\phi_q, \theta_q)$ to a HRTF $H(\phi_d, \theta_d)$ can be performed by (order-limited) SH transform at an order N [Eq. (A1)] and inverse SH transform [Eq. (A2)],

https://doi.org/10.1121/10.0009277

$$H_{nm}(\omega) = \int_{\Omega} H(\phi_q, \theta_q, \omega) Y_n^m(\theta, \phi)^* \mathrm{d}A_{\Omega}, \tag{A1}$$

$$H(\phi_d, \theta_d, \omega) = \sum_{n=0}^{N} \sum_{m=-n}^{n} H_{nm}(\omega) Y_n^m(\theta_d, \phi_d). \tag{A2}$$

Substituting the plane wave density function calculated with Eq. (2) into the binaural reproduction described with Eq. (3) yields

$$B(\omega) = \frac{1}{4\pi} \int_{\Omega} H(\phi, \theta, \omega) \times \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \frac{1}{\mathrm{i}^n j_n\left(\frac{\omega}{c} r_0\right)} S_{nm}(\omega) Y_n^m(\phi, \theta) \mathrm{d}A_{\Omega}. \tag{A3}$$

According to Williams (1999, p. 259), the SH coefficients of a unity plane wave impinging from (ϕ_d, θ_d) are

$$\tilde{S}_{nm}(\omega) = 4\pi \mathrm{i}^n j_n\left(\frac{\omega}{c} r_0\right) Y_n^m(\phi_d, \theta_d)^*. \tag{A4}$$

Inserting $\tilde{S}_{nm}$ for the sound field S_{nm} in Eq. (A3) yields

$$B(\omega) = \int_{\Omega} H(\phi, \theta, \omega) \sum_{n=0}^{\infty} \sum_{m=-n}^{n} Y_n^m(\phi_d, \theta_d)^* Y_n^m(\phi, \theta) \mathrm{d}A_{\Omega}. \tag{A5}$$

The HRTFs can be expressed as the SH sum $H_{nm}(\omega)$,

$$B(\omega) = \int_{\Omega} \sum_{n=0}^{N} \sum_{m=-n}^{n} H_{nm}(\omega) Y_n^m(\theta, \phi) \times \sum_{n=0}^{\infty} \sum_{m=-n}^{n} Y_n^m(\phi_d, \theta_d)^* Y_n^m(\phi, \theta) \mathrm{d}A_{\Omega}, \tag{A6}$$

assuming that there are no undersampling effects resulting from the discrete sampling of the sound field. Hence, the orthogonality property of the SH function holds such that

$$B(\omega) = \int_{\Omega} \sum_{n=0}^{N} \sum_{m=-n}^{n} H_{nm}(\omega) Y_n^m(\theta, \phi) \times \delta(\phi - \phi_d) \delta(\cos(\theta - \cos\,\theta_d)) \mathrm{d}A_{\Omega}. \tag{A7}$$

Resolving the integral leads to

$$H(\phi_d, \theta_d, \omega) = \sum_{n=0}^{N} \sum_{m=-n}^{n} H_{nm}(\omega) Y_n^m(\theta_d, \phi_d), \tag{A8}$$

which is exactly Eq. (A2). Hence, the binaural signals in both of the experiments are affected by exactly the same artifacts due to the order-limited SH processing. The signals in experiment 1 are additionally impaired by the undersampling artifacts because of the sampling with the SMA.

TABLE II. The results of the mixed $4 \times 3 \times 2 \times 2$ ANOVA with the between-subject factor room (R) and the within-subject factors BRIR component (B), position (P), and signal (S) for experiment 1.

Effect	Degrees of freedom (df)	F	p[a]	ε[b]	η_G^2[c]
R	3,32	3.38	0.030*	1.0	0.064
B	2,64	266.01	<0.001*	0.75	0.066
P	1,32	0.05	0.817	1.0	0.00
S	1,32	0.0	1.0	1.0	0.00
$R \times B$	6,64	7.5	<0.001*	0.75	0.14
$R \times P$	3,32	3.66	0.022*	1.0	0.016
$R \times S$	3,32	8.51	<0.001*	1.0	0.073
$B \times P$	2,64	2.73	0.078	0.93	0.01
$B \times S$	2,64	19.12	<0.001*	0.94	0.07
$P \times S$	1,32	2.17	0.150	1.0	0.003
$R \times B \times P$	6,64	1.35	0.254	0.93	0.014
$R \times B \times S$	6,64	3.13	0.011*	0.94	0.035
$R \times \mathrm{P} \times S$	3,32	2.23	0.104	1.0	0.01
$B \times P \times S$	2,64	1.37	0.261	0.89	0.005
$R \times B \times P \times S$	6,64	0.8	0.560	0.89	0.009

[a]p, The Greenhouse-Geisser corrected p-values with the statistical significance at the 5% level as indicated by the asterisks.
[b]ε, the Greenhouse-Geisser epsilons (note that only the factors with more than one level can be corrected for the sphericity).
[c]η_G^2, the generalized eta squared according to Olejnik and Algina (2003).

TABLE III. The results of the nested $3 \times 2 \times 2$ ANOVA with the within-subject factors BRIR component (B), position (P), and signal (S) for the data of room CR1 are shown for experiment 1.

Effect	df	F	p[a]	ε[b]	η_G^2[c]
B	2,16	90.905	<0.001*	0.744	0.772
P	1,8	0.175	0.687	1	0.0
S	1,8	0.006	0.939	1	0.0
$B \times P$	2,16	2.120	0.159	0.893	0.023
$B \times S$	2,16	4.485	0.039	0.814	0.04
$P \times S$	1,8	4.392	0.069	1	0.014
$B \times P \times C$	2,16	1.196	0.321	0.75	0.031

[a]p, The Greenhouse-Geisser corrected p-values and the statistical significance at the 5% level as indicated by the asterisks.
[b]ε, the Greenhouse-Geisser epsilons (note that only the factors with more than one level can be corrected for the sphericity).
[c]η_G^2, the generalized eta squared according to Olejnik and Algina (2003).

TABLE IV. The results of the nested $3 \times 2 \times 2$ ANOVA with the within-subject factors BRIR component (B), position (P), and signal (S) for the data of room CR7 are shown for experiment 1.

Effect	df	F	p[a]	ε[b]	η_G^2[c]
B	2,16	59.115	0.000	0.906	0.683
P	1,8	8.902	0.018	1	0.041
S	1,8	30.921	0.001	1	0.205
$B \times P$	2,16	0.632	0.541	0.975	0.012
$B \times S$	2,16	20.20	0.001	0.626	0.288
$P \times S$	1,8	2.778	0.134	1	0.021
$B \times P \times S$	2,16	0.121	0.850	0.825	0.001

[a]p, The Greenhouse-Geisser corrected p-values with statistical significance at the 5% level as indicated by asterisks.
[b]ε, the Greenhouse-Geisser epsilons (note that only the factors with more than one level can be corrected for the sphericity.
[c]η_G^2, the generalized eta squared according to Olejnik and Algina (2003).

TABLE V. The results of the nested $3 \times 2 \times 2$ ANOVA with the within-subject factors BRIR component (B), position (P), and signal (S) for the data of room SBS are shown for experiment 1.

Effect	df	F	p[a]	ε[b]	η_G^2[c]
B	2,16	39.642	0.000	0.573	0.549
P	1,8	2.200	0.176	1	0.019
S	1,8	1.408	0.269	1	0.017
$B \times P$	2,16	2.447	0.137	0.750	0.042
$B \times S$	2,16	10.083	0.003	0.832	0.062
$P \times S$	1,8	0.805	0.396	1	0.004
$B \times P \times S$	2,16	2.186	0.151	0.907	0.018

[a]p, The Greenhouse-Geisser corrected p-values with statistical significance at the 5% level as indicated by asterisks.
[b]ε, the Greenhouse-Geisser epsilons (note that only the factors with more than one level can be corrected for the sphericity).
[c]η_G^2, the generalized eta squared according to Olejnik and Algina (2003).

APPENDIX B: ADDITIONAL INFORMATION OF THE APPLIED STATISTICS

This appendix displays additional information of the applied statistics presented in Secs. III and IV.[2]

1. Experiment 1

Olejnik and Algina (2003) proposed the generalized eta squared as a measure for the effect size in repeated measures ANOVAs. This was supported by Bakeman (2005). Therefore, we report the generalized eta squared in ANOVA Tables II and III.

a. t-test results

1. Pairwise t-tests between rooms with the pooled data of signal, position, and BRIR component

(1) CR1 vs CR7: $(t(107) = 3.068, p < 0.055, d = 0.246)$
(2) CR1 vs SBS: $(t(107) = 1.121, p < 1.000, d = 0.100)$
(3) CR1 vs LBS: $(t(107) = 0.599, p < 1.000, d = 0.049)$
(4) CR7 vs SBS: $(t(107) = 3.991, p < 0.002^*, d = 0.391)$
(5) CR7 vs LBS: $(t(107) = 3.458, p < 0.016^*, d = 0.338)$
(6) SBS vs LBS: $(t(107) = 0.701, p < 1.000, d = 0.059)$

TABLE VI. The results of the nested $3 \times 2 \times 2$ ANOVA with the within-subject factors BRIR component (B), position (P), and signal (S) for the data of room LBS are shown for experiment 1.

Effect	df	F	p[a]	ε[b]	η_G^2[c]
B	2,16	143.959	0.000	0.661	0.66
P	1,8	1.087	0.328	1	0.007
S	1,8	6.503	0.034	1	093
$B \times P$	2,16	0.939	0.402	0.865	0.008
$B \times S$	2,16	0.287	0.691	0.741	0.01
$P \times S$	1,8	1.800	0.217	1	0.012
$B \times P \times S$	2,16	0.020	0.966	0.834	000

[a]p, The Greenhouse-Geisser corrected p-values with statistical significance at the 5% level as indicated by the asterisks.
[b]ε, the Greenhouse-Geisser epsilons (note that only the factors with more than one level can be corrected for the sphericity).
[c]η_G^2, the generalized eta squared according to Olejnik and Algina (2003).

TABLE VII. The results of the nested $4 \times 2 \times 2$ ANOVA with the between-subject factor room and within-subject factors position (P) and signal (S) for the data of the direct sound are shown for experiment 1.

Effect	df	F	p[a]	η_G^2[b]
R	3,32	0.569	0.639	0.0280
P	1,32	3.046	0.091	0.013
S	1,32	8.364	0.007	0.034
$R \times P$	3,32	2.716	0.061	0.053
$R \times S$	3,32	0.394	0.758	0.008
$P \times S$	1,32	0.285	0.597	0.001
$R \times P \times S$	3,32	0.855	0.474	0.01

[a]p, The Greenhouse-Geisser corrected p-values with statistical significance at the 5% level as indicated by the asterisks.
[b]η_G^2, the generalized eta squared according to Olejnik and Algina (2003).

2. t-tests between early reflections and reverberation part for SBS and LBS

(1) SBS $(t(35) = 3.466, p < 0.028^*, d = 0.79)$
(2) LBS $(t(35) = 3.651, p < 0.017^*, d = 0.762)$

3. Pairwise t-tests between signal 1 and signal 2 for each room and BRIR component separately

(1) CR1 DS − signal 1 vs signal 2: $(t(17) = 1.714, p < 1.000, d = 0.416)$
(2) CR1 ER − signal 1 vs signal 2: $(t(17) = 0.788, p < 1.000, d = 0.282)$
(3) CR1 REV − signal 1 vs signal 2: $(t(17) = -1.118, p < 1.000, d = 0.436)$
(4) CR7 DS − signal 1 vs signal 2: $(t(17) = 1.092, p < 1.000, d = 0.264)$
(5) CR7 ER − signal 1 vs signal 2: $(t(107) = 0.599, p < 1.000, d = 0.593)$
(6) CR7 Rev − signal 1 vs signal 2: $(t(17) = 10.547, p < 0.001, d = 2.451)$
(7) SBS DS − signal 1 vs signal 2: $(t(17) = 2.496, p < 0.463, d = 0.445)$
(8) SBS ER − signal 1 vs signal 2: $(t(17) = 2.268, p < 0.733, d = 0.722)$
(9) SBS Rev − signal 1 vs signal 2: $(t(17) = 1.916, p < 1.000, d = 0.671)$

TABLE VIII. The determined PSEs averaged across subjects and rooms with respect to BRIR component, source position, and test signal are shown for experiment 1; additionally, the 95% between-subject confidence intervals are presented. The room had a significant effect, which is why the interpretation of the mean values should be performed with reservation.

		DS	ER	REV
		PSE ± CI	PSE ± CI	PSE ± CI
Noise	30°	19 ± 1.3	12 ± 1.0	7 ± 1.0
	270°	20 ± 1.6	13 ± 1.16	6 ± 0.96
Speech	30°	17 ± 1.3	13 ± 1.35	9 ± 1.4
	270°	18 ± 1.44	12 ± 1.15	9 ± 1.2

https://doi.org/10.1121/10.0009277

TABLE IX. The results of the mixed $4 \times 3 \times 2 \times 2$ ANOVA with the between-subject factor room (R) and the within-subject factors BRIR component (B), position (P), and signal (S) are shown for experiment 2.

Effect	df	F	p[a]	ε[b]	η_G^2[c]
R	3,32	2.41	0.086	1.0	0.039
B	2, 64	122.29	$<0.001^*$	0.65	0.52
P	1, 32	6.80	0.014^*	1.0	0.012
S	1, 32	3.80	0.06	1.0	0.009
$R \times B$	6, 64	0.91	0.466	0.65	0.024
$R \times P$	3, 32	2.24	0.102	1.0	0.012
$R \times S$	3, 32	0.17	0.913	1.0	0.001
$B \times P$	2, 64	5.75	0.006^*	0.95	0.02
$B \times S$	2, 64	13.98	$<0.001^*$	0.80	0.04
$P \times S$	1, 32	0.04	0.854	1.0	0.00
$R \times B \times P$	6, 64	1.43	0.221	0.95	0.015
$R \times B \times S$	6, 64	0.96	0.447	0.8	0.009
$R \times P \times S$	3, 32	1.84	0.159	1.0	0.01
$B \times P \times S$	2, 64	1.07	0.34	0.85	0.004
$R \times B \times P \times S$	6, 64	0.54	0.75	0.85	0.007

[a]p, The Greenhouse-Geisser corrected p-values with statistical significance at the 5% level as indicated by the asterisks.
[b]ε, the Greenhouse-Geisser epsilons (note that only the factors with more than one level can be corrected for the sphericity).
[c]η_p^2, the generalized eta squared according to Olejnik and Algina (2003).

(10) LBS DS − signal 1 vs signal 2: $(t(17) = 1.648, p < 1.000, d = 0.650)$
(11) LBS ER − signal 1 vs signal 2: $(t(17) = 2.665, p < 0.327, d = 0.736)$
(12) LBS REV − signal 1 vs signal 2: $(t(17) = 1.479, p < 1.000, d = 0.489)$

2. Experiment 2

a. t-test results

1. t-tests between signal 1 vs signal 2 for each BRIR component separately

(1) DS: $(t(71) = 3.764, p < 0.003^*, d = 0.481)$
(2) ER: $(t(71) = 0.705, p < 1.000, d = 0.125)$
(3) REV: $(t(71) = 2.261, p < 0.242, d = 0.311)$

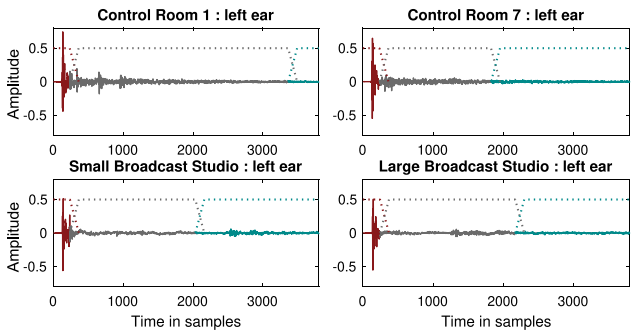


FIG. 10. (Color online) The different BRIR components and the corresponding fading windows at the transition points, as an example, for the left ear signal and each room examined are depicted.

TABLE X. The determined PSEs averaged across subjects and rooms for the conditions BRIR component, source positon, and test signal separately are shown for experiment 2; additionally, the 95% between-subject confidence intervals are presented.

		DS	ER	REV
		PSE ± CI	PSE ± CI	PSE ± CI
Noise	30°	11 ± 1.67	4 ± 0.73	4 ± 0.63
	270°	13 ± 1.66	5 ± 0.70	4 ± 0.79
Speech	30°	9 ± 1.4	5 ± 0.6	5 ± 1.0
	270°	11 ± 1.14	4 ± 0.82	5 ± 0.75

2. t-tests between position 1 vs position 2 for each BRIR component separately

(1) DS: $(t(71) = 3.240, p < 0.016^*, d = 0.403)$
(2) ER: $(t(71) = 0.587, p < 1.000, d = 0.098)$
(3) REV: $(t(71) = 0.177, p < 1.000, d = 0.029)$

3. Pairwise t-tests between all BRIR components with the pooled data of room, signal, and position

(1) DS vs ER $(t(143) = 15.986, p < 0.001^*, d = 1.796)$
(2) DS vs REV $(t(143) = 15.488, p < 0.001^*, d = 1.796)$
(3) ER vs Rev $(t(143) = 0.609, p < 1.000, d = 0.068)$

APPENDIX C: BRIR COMPONENTS AND FADING WINDOWS

Supplementary to Table I, Fig. 10 shows the left ear BRIRs for the source to the left (such that the left ear is ipsilateral) for each room and the 29th-order reference BRIR. Additionally, the linear fades are marked as dashed lines. The linear fade was performed over 128 samples so that the last 64 samples of the corresponding BRIR component were faded in and out, respectively. The employed Neumann KU100 HRIRs have a length of 128 samples. The direct sound component was defined as the first 3.5 ms (168 samples) after the onset. The mixing times were calculated for the frontal direction, which were averaged over the left and right ears according to Abel and Huang (2006) with *AKmixingTimeAbel* from the MATLAB toolbox AKtools (Brinkmann and Weinzierl, 2017).[3]

[1]See https://doi.org/10.5281/zenodo.5862771 for a detailed MATLAB code to generate Fig. 1 (Last viewed 1/22/2021).
[2]See https://doi.org/10.5281/zenodo.5862771 for a full data set of statistical results in.mat and.R format, as well as R scripts for the presented statistical analysis (Last viewed 1/22/2021).
[3]See https://doi.org/10.5281/zenodo.5862771 for a detailed MATLAB code to generate Fig. 10 (Last viewed 1/22/2021).

Abel, J. S., and Huang, P. (**2006**). "A simple, robust measure of reverberation echo density," in *Proceedings of 121th AES Convention*.
Ahrens, J. (**2019**). "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*, New Paltz, NY, pp. 2–6.
Ahrens, J., and Andersson, C. (**2019**). "Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre," J. Acoust. Soc. Am. **145**, 2783–2794.

Ahrens, J., Hohnerlein, C., and Andersson, C. (**2017**). "Authentic auralization of acoustic spaces based on spherical microphone array recordings," in *ASA/EAA Meeting*, Boston, Vol. 40, pp. 303–310.

Arend, J. M., Brinkmann, F., and Pörschmann, C. (**2021**). "Assessing spherical harmonics interpolation of time-aligned head-related transfer functions," J. Audio Eng. Soc. **69**(1/2), 104–117.

Aussal, M., Alouges, F., and Katz, B. (**2013**). "A study of spherical harmonics interpolation for HRTF exchange," in *Proceedings of Meetings on Acoustics*, Montreal, Canada, Vol. 19.

Bakeman, R. (**2005**). "Recommended effect size statistics for repeated measures designs," Behav. Res. Methods **37**(3), 379–384.

Barron, M. (**1971**). "The subjective effects of first reflections in concert halls-The need for lateral reflections," J. Sound Vib. **15**(4), 475–494.

Bee Wah, Y., and Mohd Razali, N. (**2011**). "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests," J. Stat. Model. Anal. **2**, 21–33.

Ben-Hur, Z., Alon, D. L., Mehra, R., and Rafaely, B. (**2019a**). "Efficient representation and sparse sampling of head-related transfer functions using phase-correction based on ear alignment," in *Proceedings of the IEEE/ACM Transactions on Audio Speech and Language Processing*, Vol. 27, pp. 2249–2262.

Ben-Hur, Z., Alon, D. L., Rafaely, B., and Mehra, R. (**2019b**). "Loudness stability of binaural sound with spherical harmonic representation of sparse head-related transfer functions," EURASIP J. Audio, Speech, Music Process. **2019**(1), 1–14.

Ben-Hur, Z., Sheaffer, J., and Rafaely, B. (**2018**). "Joint sampling theory and subjective investigation of plane-wave and spherical harmonics formulations for binaural reproduction," Appl. Acoust. **134**, 138–144.

Bernschütz, B. (**2013**). "A spherical far field HRIR/HRTF compilation of the Neumann KU 100," in *Proceedings of the 39th DAGA*, Meran, pp. 592–595.

Bernschütz, B. (**2016**). "Microphone arrays and sound field decomposition for dynamic binaural recording," Ph.D. thesis, Technische Universität Berlin, available at https://doi.org/10.14279/depositonce-5082 (Last viewed 1/22/2021).

Bernschütz, B., Giner, A. V., Pörschmann, C., and Arend, J. M. (**2014**). "Binaural reproduction of plane waves with reduced modal order," Acta Acust. Acust. **100**(5), 972–983.

Bernschütz, B., Pörschmann, C., Spors, S., and Weinzierl, S. (**2010**). "Entwurf und Aufbau eines variablen sphärischen Mikrofonarrays für Forschungsanwendungen in Raumakustik und Virtual Audio" (Design and construction of a variable spherical microphone array for research applications in room acoustics and virtual audio.), in *Proceedings of 36th DAGA*, pp. 717–718.

Bernschütz, B., Pörschmann, C., Spors, S., and Weinzierl, S. (**2011a**). "SOFiA Sound Field Analysis Toolbox," in *Proceedings of the International Conference on Spatial Audio (ICSA)*, pp. 8–16.

Bernschütz, B., Pörschmann, C., Spors, S., and Weinzierl, S. (**2011b**). "Soft-Limiting der modalen Amplitudenverstärkung bei sphärischen Mikrofonarrays im Plane Wave Decomposition Verfahren" ("■"), in *Proceedings of the 37th DAGA*, 2, Düsseldorf, pp. 661–662.

Blauert, J. (**1996**). *Spatial Hearing* (Hirzel, Stuttgart), p. 459.

Bortz, J., and Schuster, C. (**2010**). *Statistik Für Human- Und Sozialwissenschaftler* Statistics For Human And Social Scientists, 7th ed. (Springer, Gießen), pp. 117–136.

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

Brinkmann, F., Lindau, A., Vrhovnik, M., and Weinzierl, S. (**2014**). "Assessing the authenticity of individual dynamic binaural synthesis," in *EAA Joint Symposium on Auralization and Ambisonics*, April, Vol. 71, pp. 3–5, available at https://depositonce.tu-berlin.de/handle/11303/168 (Last viewed 1/22/2021).

Brinkmann, F., and Weinzierl, S. (**2017**). "Aktools—An open software toolbox for signal acquisition, processing, and inspection in acoustics," in *Proceedings of the 142nd AES Convention*, AES, Berlin, Germany, pp. 1–6.

Chen, S. Y., Feng, Z., and Yi, X. (**2017**). "A general introduction to adjustment for multiple comparisons," J. Thorac. Dis. **9**(6), 1725–1729.

Engel, I., Henry, C., Garí, S. V. A., Robinson, P. W., and Picinali, L. (**2021**). "Perceptual implications of different Ambisonics-based methods for binaural reverberation," J. Acoust. Soc. Am. **149**, 895.

Engel, I., Henry, C., Garí, S. V. A., Robinson, P. W., Poirier-Quinot, D., and Picinali, L. (**2019**). "Perceptual comparison of Ambisonics-based reverberation methods in binaural listening," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, Paris pp. 121–126.

Erbes, V., Wierstorf, H., Geier, M., and Spors, S. (**2017**). "Free database of low-frequency corrected head-related transfer functions and headphone compensation filter," in *Proceedings of the 142nd AES Convention*, pp. 1–5.

Erdfelder, E., Faul, F., Buchner, A., and Lang, A. G. (**2009**). "Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses," Behav. Res. Methods **41**(4), 1149–1160.

Garí, S. V. A., Brimijoin, W. O., Hassager, H. G., and Robinson, P. W. (**2019**). "Flexible binaural resynthesis of roomi responses for augmented reality research," in *Spatial AudioSignal Processing Symposium*, pp. 161–166, available at https://hal.archives-ouvertes.fr/hal-02275193/document (Last viewed 1/22/2021).

Geier, M., Ahrens, J., and Spors, S. (**2008**). "The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods," in *Proceedings of the 124th AES Convention*, Amsterdam, NE, pp. 179–184.

Geier, M., Ahrens, J., and Spors, S. (**2019**). "The SoundScape Renderer," available at http://spatialaudio.net/ssr/ (Last viewed 1/22/2021).

Giner, A. V. (**2013**). "Scale—A software tool for listening experiments," in *Proceedings of the 39th DAGA*, pp. 1–4.

Girden, E. R. (**1992**). *ANOVA: Repeated Measures* (Sage, Newbury Park, CA).

Jarmasz, J., and Hollands, J. G. (**2009**). "Confidence intervals in repeated-measures designs: The number of observations principle," Can. J. Exp. Psychol. **63**(2), 124–138.

Jeong, C. H. (**2016**). "Diffuse sound field: Challenges and misconceptions," in *Proceedings of the 45th International Congress and Exposition on Noise Control Engineering: Towards a Quieter Future*, Vol. 4, 1015–1021.

Jot, J.-M., Larcher, V., and Pernaux, J.-M. (**1999**). "A comparative study of 3-D audio encoding and rendering techniques," in *AES 16th International Conference*, pp. 281–300.

Kingdom, F. A., and Prins, N. (**2010**). *Psychophysics: A Practical Introduction*, 1st ed. (Academic, London, UK).

Kuttruff, H. (**1973**). *Room Acoustics*, 4th ed. (Spon, London, UK).

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**(2B), 467–477.

Lindau, A. (**2014**). "Binaural resynthesis of acoustical environments—Technology and perceptual evaluation," Ph.D. thesis, pp. 1–279, available at https://depositonce.tu-berlin.de/handle/11303/4382 (Last viewed 1/22/2021).

Lindau, A., Kasanke, L., and Weinzierl, S. (**2010**). "Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses," in *Proceedings of the 128th AES Convention*.

Lindau, A., Maempel, H., and Weinzierl, S. (**2008**). "Minimum BRIR grid resolution for dynamic binaural synthesis," in *Proceedings of Acoustics 2008*, Paris, Vol. 123, pp. 3498–3498.

Loftus, G. R. (**1994**). "Using confidence intervals in within-subject designs," Psychon. Bull. Rev. **1**(4), 476–490.

Lübeck, T., Helmholz, H., Arend, J. M., Pörschmann, C., and Ahrens, J. (**2020a**). "Perceptual evaluation of mitigation approaches of impairments due to spatial undersampling in binaural rendering of spherical microphone array data: Dry acoustic environments," in *Proceedings of the International Conference on Digital Audio Effects 2020*, Vienna, Vol. 68, pp. 428–440.

Lübeck, T., Pörschmann, C., and Arend, J. M. (**2020b**). "Perception of direct sound, early reflections, and reverberation in auralizations of sparsely measured binaural room impulse responses," in *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality*, Redmond, WA, pp. 1–10.

McCormack, L., Pulkki, V., and Marschall, M. (**2020**). "Higher-order spatial impulse response rendering: Investigating the perceived effects of spherical order, dedicated diffuse rendering," J. Audio Eng. Soc. **68**(5), 338–354.

Meese, T. (**1995**). "Using the standard staircase to measure the point ofsubjective equality: A guide based on computer simulations," Percept. Psychophys. **25**(1), 16–18.

Merimaa, J., and Pulkki, V. (**2004**). "Spatial impulse response rendering," in *Proceedings of the 7th International Conference on Digital Audio Effects*, Naples, pp. 139–144.

Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B. (**1995**). "Head-related transfer functions of human subjects," J. Audio Eng. Soc. **43**(5), 300–321.
Olejnik, S., and Algina, J. (**2003**). "Generalized eta and omega squared statistics: Measures of effect size for some common research designs," Psychol. Methods **8**(4), 434–447.
Olive, S. E., and Toole, F. E. (**1988**). "The detection of reflections in typical rooms," in *AES 85th Convention*, Ottawa, Canada, Vol. 2719.
Pearson, E. S. (**1931**). "The analysis of variance in cases of non-normal variation," Biometrika **23**(1/2), pp. 114–133.
Pike, C. W. (**2019**). "Evaluating the perceived quality of binaural technology," Ph.D. thesis, University of York.
Pörschmann, C., Arend, J. M., Bau, D., and Lübeck, T. (**2020**). "Comparison of spherical harmonics and nearest-neighbor based interpolation of head-related transfer functions," in *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality*, Redmond, WA, pp. 1–10.
Pörschmann, C., Arend, J. M., and Brinkmann, F. (**2019**). "Directional equalization of sparse head-related transfer function sets for spatial upsampling," IEEE/ACM Trans. Audio, Speech, Lang. Process. **27**(6), 1060–1071.
Pulkki, V. (**2007**). "Spatial sound reproduction with directional audio coding," J. Audio Eng. Soc. **55**(6), 503–516.
Rafaely, B. (**2015**). *Fundamentals of Spherical Array Processing* (Springer, Berlin).
Savioja, L., and Svensson, U. P. (**2015**). "Overview of geometrical room acoustic modeling techniques," J. Acoust. Soc. Am. **138**(2), 708–730.
Stade, P., Bernschütz, B., and Rühl, M. (**2012**). "A spatial audio impulse response compilation captured at the WDR Broadcast Studios," in *Proceedings of the 27th Tonmeistertagung—VDT International Convention*, pp. 551–567.
Tervo, S., Pätynen, J., Kuusinen, A., and Lokki, T. (**2013**). "Spatial decomposition method for room impulse responses," J. Audio Eng. Soc. **61**(1/2), 17–28.
Vorländer, M. (**2008**). *Auralization*, 1st ed. (Springer, Berlin), pp. 1–335.
Williams, E. G. (**1999**). *Fourier Acoustics* (Academic, London), p. 302.
Yap, B. W., and Sim, C. H. (**2011**). "Comparisons of various types of normality tests," J. Stat. Comput. Simul. **81**(12), 2141–2155.
Zahorik, P. (**2002**). "Direct-to-reverberant energy ratio sensitivity," J. Acoust. Soc. Am. **112**(5), 2110–2117.
Zaunschirm, M., Schörkhuber, C., and Höldrich, R. (**2018**). "Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," J. Acoust. Soc. Am. **143**(6), 3616–3627.
Zotter, F. (**2009**). "Analysis and synthesis of sound-radiation with spherical arrays," Ph.D. thesis, University of Music and Performing Arts, Austria.

5.2 Efficient Binaural Rendering of Spherical Microphone Array Data by Linear Filtering

Arend*, J. M., Lübeck*, T., & Pörschmann*, C. (2021). *EURASIP J. Audio Speech Music Process.*, *2021*(37), 1–11. (*equal contributions). https://doi.org/10.1186/s13636-021-00224-5

Arend *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:37
https://doi.org/10.1186/s13636-021-00224-5

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Efficient binaural rendering of spherical microphone array data by linear filtering

Johannes M. Arend[1,2*†], Tim Lübeck[1,2†] and Christoph Pörschmann[1†]

Abstract

High-quality rendering of spatial sound fields in real-time is becoming increasingly important with the steadily growing interest in virtual and augmented reality technologies. Typically, a spherical microphone array (SMA) is used to capture a spatial sound field. The captured sound field can be reproduced over headphones in real-time using binaural rendering, virtually placing a single listener in the sound field. Common methods for binaural rendering first spatially encode the sound field by transforming it to the spherical harmonics domain and then decode the sound field binaurally by combining it with head-related transfer functions (HRTFs). However, these rendering methods are computationally demanding, especially for high-order SMAs, and require implementing quite sophisticated real-time signal processing. This paper presents a computationally more efficient method for real-time binaural rendering of SMA signals by linear filtering. The proposed method allows representing any common rendering chain as a set of precomputed finite impulse response filters, which are then applied to the SMA signals in real-time using fast convolution to produce the binaural signals. Results of the technical evaluation show that the presented approach is equivalent to conventional rendering methods while being computationally less demanding and easier to implement using any real-time convolution system. However, the lower computational complexity goes along with lower flexibility. On the one hand, encoding and decoding are no longer decoupled, and on the other hand, sound field transformations in the SH domain can no longer be performed. Consequently, in the proposed method, a filter set must be precomputed and stored for each possible head orientation of the listener, leading to higher memory requirements than the conventional methods. As such, the approach is particularly well suited for efficient real-time binaural rendering of SMA signals in a fixed setup where usually a limited range of head orientations is sufficient, such as live concert streaming or VR teleconferencing.

Keywords: Spherical microphone arrays, Binaural rendering, Spatial audio reproduction, Virtual acoustics

1 Introduction

Headphone-based binaural rendering of spatial sound fields is of great importance in the consumer sector for virtual reality (VR) and augmented reality (AR) applications as well as in research areas such as hearing science. Using a spherical microphone array (SMA) is a flexible method to capture a spatial sound field and render it for a single listener over headphones. One possibility is to measure spatial room impulse responses (SRIRs) with an SMA, which can then be used to generate binaural room impulse responses (BRIRs) [1–5]. To auralize the captured sound field, dynamic binaural synthesis is employed, i.e., the generated BRIRs are convolved (in real-time) with anechoic audio material. However, the major advantage of SMAs is that they can be used for real-time rendering of a spatial sound scene, such as a musical performance in a concert hall. In this case, the captured sound field is processed in real-time to generate ear signals that, when presented over headphones, virtually place the listener in the sound

*Correspondence: Johannes.Arend@th-koeln.de
†Johannes M. Arend, Tim Lübeck and Christoph Pörschmann contributed equally to this work.
[1]Institute of Communications Engineering, TH Köln - University of Applied Sciences, Betzdorfer Str. 2, 50679, Cologne, Germany
[2]Audio Communication Group, Technical University of Berlin, Einsteinufer 17c, 10587, Berlin, Germany

field [6–8]. Both methods allow binaural rendering with head tracking, i.e., rendering for arbitrary head orientations of the listener. Furthermore, individual head-related transfer functions (HRTFs) can be employed for binaural rendering of SMA data.

Recent advances in research yielded several solutions for binaural real-time rendering of SMA signals, such as the IEM Plug-in Suite [6, 9], SPARTA [8], and ReTiSAR [7, 10]. The overall concept of these toolboxes is similar. The sound field captured with an SMA is first spatially encoded in real-time, i.e., it is transformed to the spherical harmonics (SH) domain using the discrete SH transform (SHT) [11]. The resulting SH signals are then processed with radial filters, which are array-specific filter functions that compensate for the spatial extent and, in the case of a rigid sphere array, the scattering properties of the array body [3, 4]. A classical approach for binaural decoding of SH signals (also referred to as Ambisonics signals) is the use of so-called virtual loudspeakers [2, 6, 12–14]. By applying the inverse SH transform (ISHT) to the SH signals, spatially uniformly distributed plane waves are generated, which are then weighted with HRTFs of the corresponding directions. More recent methods perform binaural rendering directly in the SH domain, i.e., the HRTF set is transformed to the SH domain and then multiplied with the SH signals of the array [6, 7]. Both rendering methods are usually combined with further pre- or postprocessing methods, such as max-$\mathbf{r}_\mathrm{E}$ weighting [15], SH tapering [16], spherical head filters [17], or MagLS [6], to mitigate spatial aliasing and truncation errors caused by spatial discretization of the sound field in SMA capturing (see [5] for an overview of different mitigation approaches).

Real-time binaural rendering of SMA signals in the manner described above is computationally demanding, in particular because of the time-consuming SHT. Due to these performance requirements, the most recent implementation of ReTiSAR, for example, can only render SMA data up to a maximum spatial order of $N = 12$ on a standard laptop [10]. This spatial order corresponds to an SMA with a minimum number of $Q = 169$ microphones ($Q = (N+1)^2$) and is thus sufficient for most common SMAs available in commercial or scientific contexts, which mostly do not exceed an SH order of $N = 7$ (e.g., em32 Eigenmike [18], Zylia ZM-1 [19], HØSMA [20]). However, content based on sequentially measured higher-order SMA data (see, for example [21] or [22], providing SMA data with $N = 29$ or $N = 44$, respectively) cannot be rendered in real-time with current implementations. Furthermore, approaches that perform spatial upsampling of real-world SMA signals before the SHT to enhance the rendering [23] significantly increase the spatial order and thus the number of audio channels, making real-time rendering of upsampled SMA signals impossible with currently available implementations.

In this paper, we present a simpler and computationally more efficient approach for real-time binaural rendering of SMA signals. As the entire encoding and decoding chain represents a linear time-invariant (LTI) system, it can also be described with a set of finite impulse response (FIR) filters. More precisely, the transmission from each microphone input to the decoded ear signal for the left or right ear can be described by one FIR filter each, resulting in a set of $Q \times 2$ FIR filters required for binaural decoding of SMA signals for one specific head orientation. Those filters can be precomputed (for any desired number of head orientations) for the specific SMA and HRTF configuration and applied to the SMA signals in real-time by fast convolution, similar to dynamic binaural synthesis. Superimposing the output of all filtered SMA signals yields exactly the ear signals produced by any of the conventional encoding and decoding chains described above, given that the settings are the same as for the FIR filter precomputation. According to their functionality to describe the transmission from the array microphones to the ears, we have named these filters SMATBIN (Spherical Microphone Array To Binaural) filters.

The proposed approach significantly reduces the complexity of implementing real-time binaural rendering of SMA signals while also being less computationally demanding. Thus, any existing software or hardware structure for efficient and fast real-time convolution can be used for binaural rendering of SMA signals of a very high order, i.e., with many audio channels. In the following, we first explain the common encoding and decoding chains briefly discussed above in greater detail. We then provide further details on the SMATBIN filter method and explain how the filters can be generated. Next, we compare BRIRs resulting from applying the SMATBIN filters with those resulting from common binaural rendering in two working examples. Finally, we compare the computational complexity and the memory requirements of the proposed approach to that of the common rendering methods and discuss the advantages and disadvantages of using the proposed filters for binaural rendering of SMA signals.

2 Binaural rendering methods

2.1 Virtual loudspeaker approach

The block diagrams in Fig. 1 show two common methods for binaural rendering of SMA data (top and middle) and the proposed approach using the SMATBIN filters (bottom). The block diagram on the top illustrates the classical virtual loudspeaker approach [2, 6, 12–14]. The Q microphone signals captured with an SMA are transformed to the SH domain employing the SHT. The resulting SH

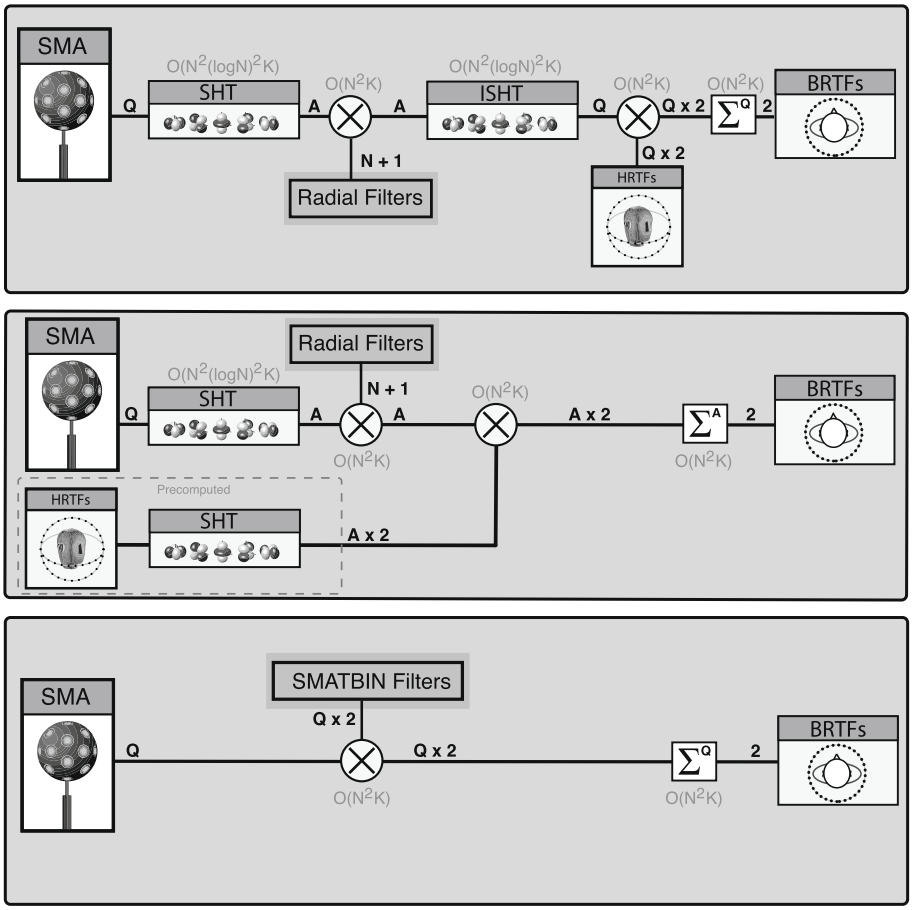


Fig. 1 Block diagrams illustrating the signal processing in the temporal frequency domain for binaural rendering of SMA data using the virtual loudspeaker approach (top), the SH domain approach (middle), and the proposed SMATBIN filter approach (bottom). *N*: spatial order, *Q*: number of microphones, *A*: number of SH channels, *O*: Landau's symbol, *K*: FFT size (top, middle) or SMATBIN filter length (bottom)

signals with A channels are multiplied with $N+1$ order-dependent radial filters, and then plane waves for specific directions are generated by applying the ISHT. This procedure, known as plane wave decomposition, is usually performed on a spatial sampling grid of the same spatial order N as that of the SMA, but different sampling schemes (e.g., Lebedev or Gaussian schemes with the same order) yield different results [2, 3, 6, 24]. For simplicity, we assume in the present case that the sound field is decomposed to Q plane waves with the directions of the SMA sampling scheme. The Q plane waves are then multiplied with Q HRTFs for the corresponding directions, resulting in the virtual loudspeakers. Finally, the Q spatially weighted plane waves are summed up, yielding the two-channel binaural signal.

The block diagram shows the processing for a single head orientation. Binaural room transfer functions (BRTFs) for arbitrary head orientations can be generated in two different ways. One way is to use HRTFs for directions corresponding to the relative angles between head orientation and the plane wave directions [2, 3, 25]. Alternatively, the sound field can be rotated in the SH domain according to the head orientation before the ISHT [8, 14]. When using complex SH basis functions, the sound field rotation can be performed by Wigner-D weighting [26], whereas for real SH basis functions, a computationally more efficient rotation matrix obtained by recursion relations can be applied [27]. The processing, including switching the HRTFs or rotating the sound field depending on the head orientation, can be performed in real-time so that the spatial sound scene captured with the SMA can be reproduced binaurally in real-time [8].

2.2 Spherical harmonics domain approach

Alternatively, binaural decoding can be directly performed in the SH domain [6, 7, 10], as illustrated in Fig. 1 (middle). As with the virtual loudspeaker approach, the SMA signals are transformed to the SH domain, and radial filters are applied. For binaural decoding, the SH signals of the array with A channels are multiplied with

the HRTF set, which was also transformed to the SH domain at the same spatial order N, resulting in an HRTF set with A SH channels. The final BRTF is obtained by summing up all A SH channels. BRTFs for arbitrary head orientations can be achieved by rotating the sound field in the SH domain applying a rotation matrix to the SH signals [26, 27]. All processing can also be done in real-time, enabling dynamic binaural auralization of SMA data [6, 7, 10]. Compared to the virtual loudspeaker approach, calculating the ear signals directly in the SH domain is less computationally expensive because the multiplication with the SH basis functions, which is part of the ISHT, is omitted.

3 SMATBIN filter method

As both above-mentioned encoding and decoding chains represent LTI systems for which the principle of superposition holds, the transmission path from each microphone of the SMA to the left and right ear can be described by a pair of FIR filters. Such a pair of filters can be calculated by applying a unit impulse (Dirac delta) to the respective channel of the SMA, while assigning zeros to the other channels, and performing the usual encoding and decoding as described above. Applying unit impulses to each channel of the SMA successively, while always assigning zeros to the other channels, yields a set of $Q \times 2$ FIR filters – the SMATBIN filters. Algorithm 1 shows the pseudocode for generating SMATBIN filters for one head orientation using either the virtual loudspeaker approach or the SH domain approach for binaural decoding. To generate SMATBIN filters for arbitrary head orientations, rotation must be integrated at the appropriate point in the algorithm. The sound field rotation in the SH domain is implemented after the radial filtering in step 5, whereas the HRTF switching is integrated in step 8 (see also Sections 2.1 and 2.2). Notably, the proposed principle can be used to convert not only the discussed common methods, but any approach for binaural rendering of SMA data, including any of the popular mitigation approaches implemented in the rendering [5], into a set of FIR filters.

The block diagram on the bottom of Fig. 1 shows the simple structure for binaural rendering when using the SMATBIN filters. Each of the Q microphone signals is convolved with the corresponding two-channel filter and then, all Q filtered microphone signals are summed up, yielding the two-channel binaural signal. The approach thus omits the computationally expensive SHT, and real-time binaural rendering can be achieved by an efficient and fast convolution of the SMA signals with the SMATBIN filters. For dynamic binaural auralization, the filter sets are precomputed for a suitably large number of head orientations, resulting in $Q \times 2 \times M$ filters, with M the number of head orientations. In real-time rendering, the SMATBIN filters are selected and switched according to the head orientation, just as any common binaural renderer does.

Algorithm 1 Pseudocode for generating the SMATBIN filters for one head orientation using either the virtual loudspeaker approach or the SH domain approach for binaural decoding. See text for more information on how to integrate head orientations.

```
for q = 1:Q do
    Apply unit impulse to the q-th microphone
    Perform FFT
    Perform SHT
    Perform radial filtering
    if Virtual Loudspeaker Approach then
        Perform ISHT
        Multiply plane waves with HRTFs for same directions
        Sum over Q weighted plane waves
    else if SH Domain Approach then
        Weight SH signals with HRTFs
        Sum over A SH channels
    end if
    Perform IFFT
    Export q-th SMATBIN filter
end for
```

4 Results

4.1 Working examples

To evaluate the proposed method, we implemented two working examples comparing binaural rendering of SMA data using the SMATBIN filters with the rendering chain implemented in the SOFiA toolbox [28]. As all spherical microphone array processing in the present work was performed using SOFiA, and the SMATBIN filters for the two examples were based on the SOFiA rendering chain, the BRTFs/BRIRs produced by the two methods should ideally be identical.

For binaural decoding, we used the `sofia_binauralX` function, which employs the virtual loudspeaker approach in combination with HRTF switching to account for arbitrary head orientations [2, 3]. The HRTFs used in SOFiA are from a Neumann KU100 dummy head measured on a Lebedev grid with 2702 sampling points [29]. The HRTFs are transformed to the SH domain at a sufficiently high order of $N = 35$, allowing artifact-free SH interpolation to obtain HRTFs for any direction corresponding to the directions of the plane waves [3].

For both working examples, the radius of the rigid sphere array was $r = 8.75$ cm, and the radial filter gain was soft-limited to 20 dB [30]. The SMATBIN filter length was defined as $K = 2048$ taps at a sampling rate of $f_s = 48$ kHz. Figure S1 in the supplementary material (Additional file 1) shows an example of SMATBIN filters with

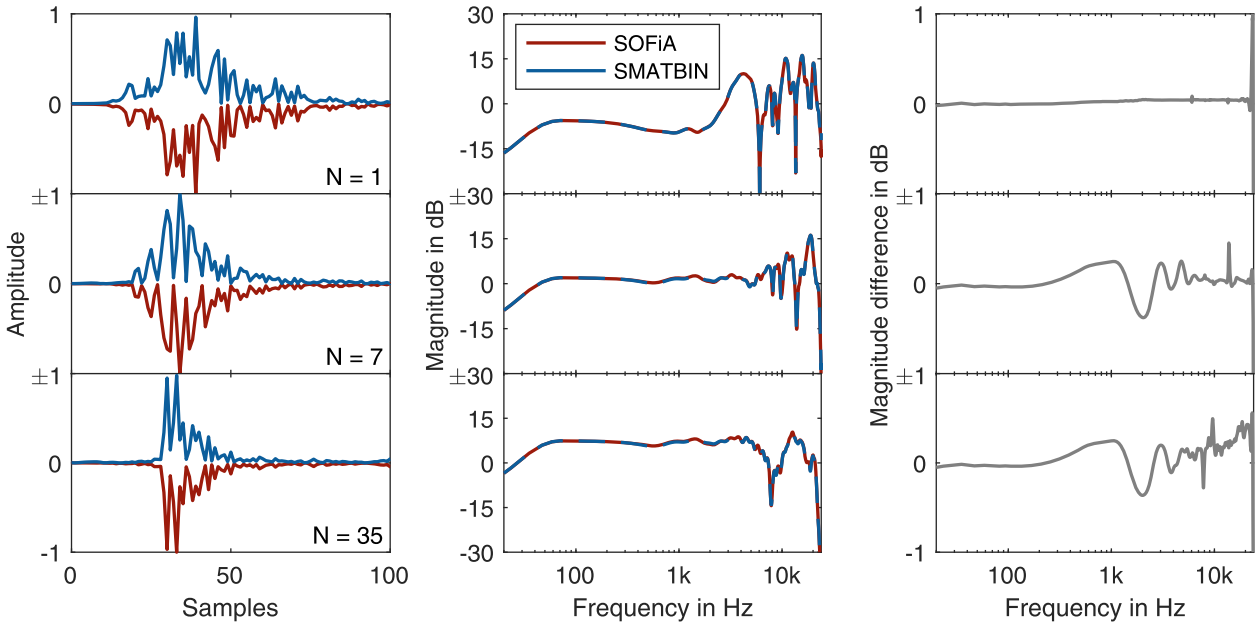


Fig. 2 Working Example 1: Left-ear BRIRs/BRTFs resulting from the SOFiA (reference) and the SMATBIN filter processing ($K = 2048$ taps) for frontal head orientation ($\phi = 0°, \theta = 90°$) and a single broadband plane wave incident from the front ($\phi = 0°, \theta = 90°$) on three different rigid sphere arrays with Lebedev sampling schemes of orders $N = \{1, 7, 35\}$. For better illustration, the absolute amplitudes of the SOFiA BRIRs in the left column are inverted. The right column shows the magnitude differences between the SOFiA and the SMATBIN BRTFs

the above-mentioned array and filter parameters for a Lebedev sampling scheme of order $N = 1$. The described implementations with functions to calculate the SMATBIN filters and generate results plots, as well as various demo implementations, are available online[1].

4.1.1 Working example 1

In the first working example, we simulated a single broadband plane wave incident from the front ($\phi = 0°$, $\theta = 90°$, with ϕ the horizontal angle ranging from 0° to 360° and θ the vertical angle ranging from 0° to 180°) on three different rigid sphere arrays with Lebedev sampling schemes of orders $N = \{1, 7, 35\}$, corresponding to 6, 86, and 1730 sampling points respectively. Besides the more common orders $N = 1$ and $N = 7$, we decided to show the implementation with the rather high order $N = 35$ to verify that no artifacts or instabilities occur even when processing with a very high number of SMATBIN filters. From these SMA signals, we calculated BRIRs using the SOFiA implementation employing plane wave decomposition and virtual loudspeaker rendering (see Fig. 1 (top)) as well as using the proposed SMATBIN filter method where the SMA signals are simply filtered and then superimposed to achieve a BRIR (see Fig. 1 (bottom)).

[1] Available: https://github.com/AudioGroupCologne/SMATBIN

Figure 2 compares the left-ear BRIRs/BRTFs resulting from the SOFiA and the SMATBIN filter processing, taking frontal head orientation ($\phi = 0°$, $\theta = 90°$) as an example. The absolute amplitudes of the broadband pressure BRIRs (left column) are nearly identical in their overall time-energy structure with matching amplitude and time events. Accordingly, the magnitude frequency responses of the respective BRTFs (middle column) show no considerable differences and are almost identical at all examined spatial orders. Consistent with that, the magnitude differences (right column) are minimal over the entire audible frequency range from 20 Hz to 20 kHz for all examined spatial orders, with a maximum of about ±0.5 dB at higher frequencies.

In further analysis, we compared BRIRs for 360 head orientations in the horizontal plane (1° steps from 0° to 360°), generated based on the SMA signals for a single plane wave incident from the front as described above. For a perception-related evaluation of the spectral deviations, we calculated for each head orientation the absolute energetic difference ΔG between SOFiA and SMATBIN BRIRs in 40 auditory gammatone filter bands between 50 Hz and 20 kHz [31, 32], as implemented in the Auditory Toolbox [33]. Figure 3 shows the so determined left-ear differences on the example of $N = 7$ for all 360 head orientations (gray lines) and averaged over all head orientations (blue line). In general, the differences are minimal and well below an assumed just-noticeable difference (JND) of 1 dB

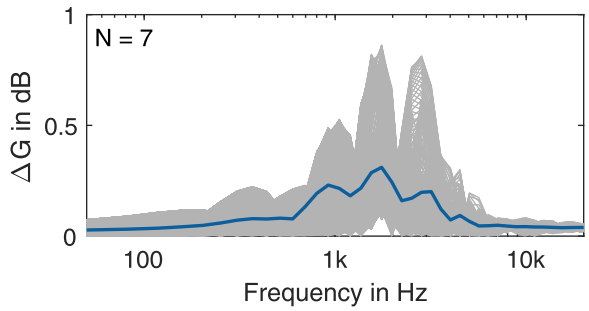


Fig. 3 Working Example 1: Left-ear absolute energetic difference ΔG in auditory gammatone filter bands between BRTFs resulting from the SOFiA (reference) and the SMATBIN filter processing ($K = 2048$ taps) for 360 head orientations in the horizontal plane (1° steps from 0° to 360°) and a single broadband plane wave incident from the front ($\phi = 0°, \theta = 90°$) on a rigid sphere array with a Lebedev sampling scheme of order $N = 7$. The gray lines show the difference for each individual head orientation, the blue line shows the difference averaged across all head orientations

[34] and thus can be considered perceptually uncritical. For certain head orientations, the differences reach a maximum of approximately 0.8 dB in the frequency range of about 2-3 kHz. These larger differences occur mainly for lateral sound incidence, i.e., for head orientations in the range of 90° and 270°. Smaller differences with a maximum of approximately 0.3 dB in the range of 2-3 kHz occur for frontal and rear sound incidence, i.e., for head orientations in the range of 0° and 180°. The average difference across head orientations is generally very small, but increases slightly towards mid frequencies, reaching a maximum of approximately 0.3 dB at about 2 kHz.

4.1.2 Working example 2

In the second working example, we evaluated the proposed method using measured SMA data of a real, more complex sound field. Specifically, we employed data captured with the VariSphear measurement system [35] on a Lebedev grid of order $N = 44$ in a classroom at TH Köln [22]. The shoebox-shaped classroom has a volume of 459m^3 and a mean reverberation time of about 0.9 s (0.5 - 8 kHz). The sound source was a Neumann KH420 loudspeaker, placed at a distance of about 4.50 m and a height of 1.40 m in front of the VariSphear array. We spatially resampled the measurements to Lebedev grids of orders $N = \{1, 7, 35\}$ using SH interpolation. From these (resampled) SMA data, we calculated BRIRs using the SOFiA rendering chain as well as the SMATBIN filter method.

Figure 4 compares the left-ear BRIRs/BRTFs for frontal head orientation generated using SOFiA or the SMATBIN filter method. Also for the complex sound field, the time-energy structure of the two broadband pressure BRIRs (left column) is almost identical. Consequently, the 1/6-octave smoothed magnitude responses (middle column) are largely identical for all spatial orders examined, and the magnitude differences (right column) are minimal, with a

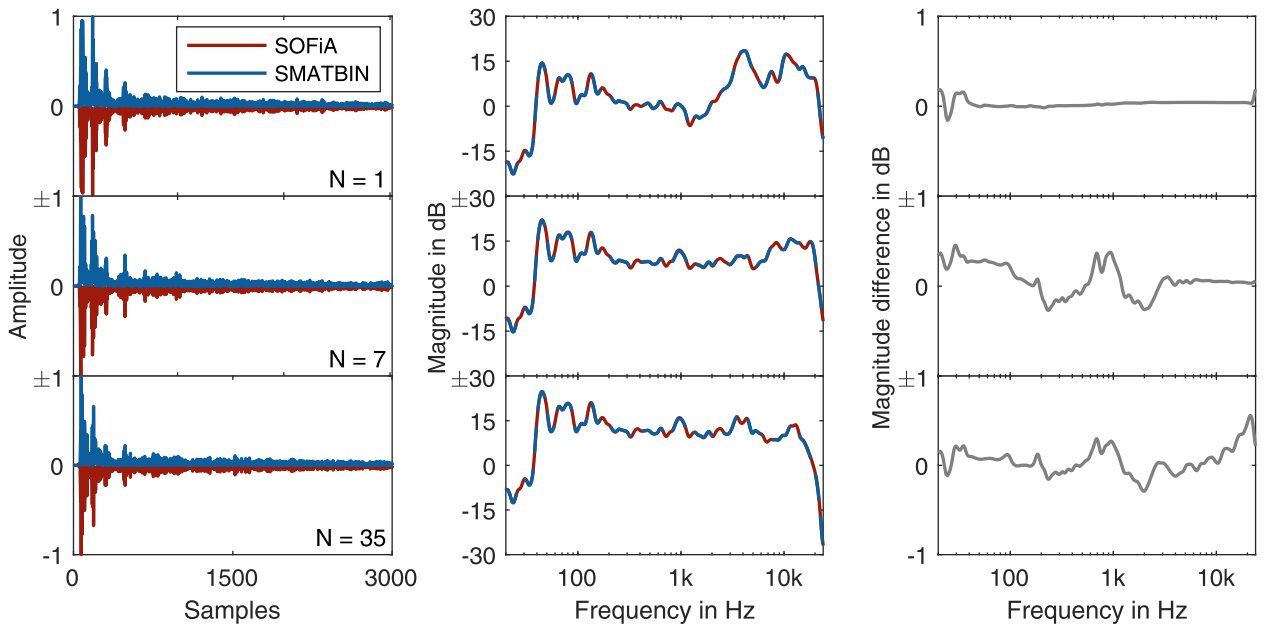


Fig. 4 Working Example 2: Left-ear BRIRs/BRTFs resulting from the SOFiA (reference) and the SMATBIN filter processing ($K = 2048$ taps) for frontal head orientation ($\phi = 0°, \theta = 90°$) and impulse responses of a classroom for three different rigid sphere arrays with Lebedev sampling schemes of orders $N = \{1, 7, 35\}$. For better illustration, the absolute amplitudes of the SOFiA BRIRs in the left column are inverted and the magnitudes of the BRTFs in the middle column are 1/6-octave smoothed. The right column shows the magnitude differences between the SOFiA and the SMATBIN BRTFs

maximum range of about ±0.5 dB over the entire audible frequency range.

The analysis of the absolute energetic difference ΔG across 360 head orientations in the horizontal plane and selected SH order $N = 7$ revealed differences that should be perceptually uncritical as they are clearly below the assumed JND of 1 dB (see Fig. 5). At frequencies below 100 Hz and in the range between 500 Hz and 3 kHz, the differences for certain head orientations reach a maximum of about 0.4 dB, but decrease again above 3 kHz. The average difference across head orientations does not exceed 0.2 dB in the entire audible frequency range and even tends towards 0 dB at frequencies above 3 kHz.

4.1.3 Interim summary

The results of the two working examples clearly show that the presented approach can be used equivalently to the established but much more complex virtual loudspeaker approach for binaural rendering of SMA data or for generating BRIRs from SMA measurements. Theoretically, the result of the two compared methods should even be completely identical. In practice, however, minimal differences between the binaural signals can occur because of the filter design, i.e., because of the necessary further processing of the filters after sampling the rendering chain with unit pulses, such as windowing, truncation, or delay compensation.

The supplementary material (Additional file 1) contains further BRIR/BRTF plots for Working example 1 for the (more application-oriented range of) orders $N = \{1, 3, 7\}$, selected SMATBIN filter lengths, and selected head orientations. Similar to Fig. 2, the results of the SOFiA and SMATBIN renderings are nearly identical as long as the SMATBIN filters have a sufficient number of filter taps. If the number of FIR filter taps is too small (approximately below 512 taps), deviations from the reference occur in the low-frequency range (<100 Hz) because of insufficient frequency resolution. The SMATBIN filter length can thus be used to adjust the accuracy of the binaural reproduction (compared to the reference) in the low-frequency range, but also the required computing power and memory requirements, as the computing effort for the real-time convolution as well as the required memory space depends on the number of filter taps.

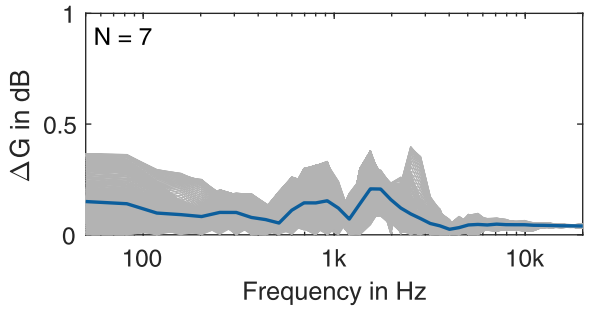


Fig. 5 Working Example 2: Left-ear absolute energetic difference ΔG in auditory gammatone filter bands between BRTFs resulting from the SOFiA (reference) and the SMATBIN filter processing ($K = 2048$ taps) for 360 head orientations in the horizontal plane (1° steps from 0° to 360°) and impulse responses of a classroom for a rigid sphere array with a Lebedev sampling scheme of order $N = 7$. The gray lines show the difference for each individual head orientation, the blue line shows the difference averaged across all head orientations

4.2 Computational complexity

In particular, towards higher orders N, the SHT dominates the computational complexity[2] of the common virtual loudspeaker and SH domain approaches. As the SHT must be performed for each frequency bin, it scales linearly with the filter length or FFT size K. The SMATBIN filter approach omits the SHT and reduces the entire encoding and decoding chain to linear filtering and summation (see Fig. 1), thereby decreasing the complexity for binaural rendering of SMA data, as detailed in the following.

The conventional SHT has a complexity of $O(N^4 K)$ and thus, the calculation effort increases rapidly as a function of spatial order N [36]. Optimized methods for performing the SHT with reduced complexity still require $O(N^2 (logN)^2 K)$ or $O(N^{\frac{5}{2}} (logN) K)$ steps, depending on the optimization [36, 37].

All other processing steps for binaural rendering of SMA data depend on N only with $O(N^2)$. The FFT and IFFT, required in all rendering methods to transform the SMA signals to frequency domain and the binaural signals to time domain, respectively, both have a complexity of $O(N^2 K logK)$. Linear filtering in frequency domain, which in the present case corresponds to either applying the radial filters to the SH signals or the SMATBIN filters to the SMA signals, has a complexity of $O(N^2 K)$, and summing up all channels also has a complexity of $O(N^2 K)$.

Thus, the SHT has the highest complexity depending on N in the entire rendering chain, and especially for large N, its calculation effort significantly exceeds that of all other processing steps. As a result, by omitting the SHT, the SMATBIN filter method allows a more efficient binaural rendering of SMA data than the conventional methods.

4.3 Memory requirements

The lower computational complexity of the SMATBIN filter method comes at the cost of higher memory requirements, as a set of filters must be precomputed and stored for each required head orientation. To estimate by example how much more memory the SMATBIN filter approach requires compared to the virtual loudspeaker or

[2] With the term computational complexity, we refer in the following to running time or time complexity.

SH domain approach, we assume in the following an SMA with a Lebedev sampling scheme of order $N = 12$, i.e., $Q = 230$ microphones and $A = 169$ SH channels, a bit depth of 32 bit, i.e., $P = 4$ bytes per filter tap, and a filter length of $K = 2048$ taps.

For the virtual loudspeaker approach, $N+1$ radial filters with a length of K taps and $2 \times D$ HRTF filters with a length of L taps must be stored, with D the number of directions of the HRTF set. The total memory requirement calculates as $[\,((N+1) \times K + 2 \times D \times L) \times P]$. Assuming an HRTF set with $D = 2702$ directions and $L = 128$ taps, the virtual loudspeaker approach requires 2.9 MB.

With the SH domain approach, only $2 \times A$ HRTF filters in the SH domain need to be stored in addition to the radial filters. Here, the total memory requirement calculates as $[\,((N+1) + A \times 2) \times K \times P]$, which also results in 2.9 MB.

In the case of the SMATBIN filter method, the required memory scales with the number of microphones Q and the number of head orientations M. The total memory requirement calculates as $[\,Q \times 2 \times M \times K \times P]$. Assuming that, as is often the case, only head orientations in the horizontal plane with a sufficiently high resolution of 2° are rendered [38], yields $M = 180$ head orientations and a total memory requirement of 678 MB. Thus, the SMATBIN filter method requires significantly more memory space than the other two methods, but is computationally less demanding. Accordingly, it must be decided on a case-by-case basis, depending on the technical requirements of a rendering system, whether memory space can be sacrificed for a lower computational load.

5 Discussion

Real-time binaural rendering of SMA data is currently being intensively researched and is becoming increasingly important for various VR and AR applications. However, common rendering methods are extremely computationally demanding, especially for high-order SMAs, and require quite sophisticated real-time signal processing. With the SMATBIN filter method, we presented in this paper a less computationally demanding approach for real-time binaural rendering of SMA data. The presented method allows representing any common rendering chain as a set of precomputed FIR filters, which are then applied to the SMA signals in real-time using fast convolution to generate the binaural signals. As the rendering process is reduced to simple linear filtering with a two-channel FIR filter per SMA channel, it is easier-to-implement using any existing hardware and software solution for fast convolution. Established binaural renderers, such as the SoundScape Renderer [39] or PyBinSim [40], are well suited for this purpose, as they already have implemented methods for optimal filter switching according to the listener's head orientation (see the demo implementation using the SoundScape Renderer in the SMATBIN repository[1]).

The technical evaluation results clearly show that the SMATBIN filer method can be used equivalently to the conventional methods. Thus, BRIRs generated with SMATBIN filtering were almost identical to BRIRs generated with the common virtual loudspeaker method [2, 3]. Furthermore, we showed that by omitting the costly SHT, rendering with SMATBIN filters has significantly lower computational complexity and is thus less computationally demanding than, for example, the virtual loudspeaker or SH domain approach [7, 10]. However, example calculations showed that the lower computational cost of the SMATBIN filter method comes along with considerably higher memory requirements than those for the virtual loudspeaker or SH domain approaches.

The advantages of lower computational complexity are not only accompanied by higher memory requirements, but also by less flexibility. As the SMATBIN filters are always precomputed for a specific SMA configuration with specific HRTFs, neither the SMA nor the HRTFs can be exchanged quickly and flexibly within an application without recalculating the filters or loading a complete precomputed filter set for the changed configuration. Moreover, the en- and decoding are no longer decoupled, and basic SH domain processing such as beamforming, sound field rotation, or spatial effects applied to the sound field in the SH domain, as available in the IEM Plug-in Suite [6, 9], are not possible at all.

Apart from our proposed method, there are alternative filtering methods for binaural rendering of microphone array captures that also omit to transform the sound field to the SH domain. One example is the virtual artificial head [41, 42], which is a filter-and-sum beamformer based on a planar microphone array with 24 microphones used to generate BRIRs. Another recent approach is beamforming-based binaural reproduction [43], with the concept of generating BRIRs directly from signals of arbitrary array structures (spherical or planar) by applying beamforming filter structures. Interestingly, depending on the parameterization of the beamformer, the results are equivalent to SH processing. For example, when using an SMA, the array output of a maximum directivity beamformer corresponds to a plane wave decomposition for the look-direction [11, 43]. Unlike the proposed filter method, however, to the best of our knowledge none of the beamformer methods have been implemented for real-time rendering of array streams, but only for generating BRIRs that are then used for auralization using dynamic binaural synthesis. That said, comparing the performance and computational demands of different beamforming-based methods with the SMATBIN filter approach in a real-time framework would be an interesting study for future research.

Although the proposed method and beamforming-based methods share some similarities in terms of using a specific filter structure for binaural rendering, sampling SH-based rendering chains as performed with the SMATBIN filter approach has some advantages. For one thing, many aspects of SH processing are well understood, both technically and perceptually, such as the required grid resolution, the frequency characteristics of the beams, or the behaviour of spatial aliasing [11], to name a few. These findings can be used to create optimized rendering chains, which can than be sampled and stored as FIR filters for more efficient binaural rendering. Furthermore, there are several approaches to mitigate undersampling errors when using real-world SMAs (e.g., max-$\mathbf{r}_\mathrm{E}$ weighting [15], SH tapering [16], spherical head filters [17], or MagLS [6]), which can also be sampled as part of the rendering chain using the SMATBIN filter method and integrated into a real-time implementation. Thus, using SMATBIN filters makes it possible to integrate any mitigation approach (under development) into a real-time framework without extensive modifications of the real-time processing chain. More specifically, any rendering chain, no matter how complex, which may be difficult to implement in real-time, can be sampled using the presented method and used for real-time rendering using a standard convolution engine.

The presented method offers advantages, particularly for fixed SMA to binaural chains that should be rendered efficiently. Due to its lower computational complexity, the SMATBIN filter method enables real-time rendering of high-order SMA signals ($N > 12$, which is currently the maximum feasible order with the real-time renderer ReTiSAR [10] on a standard laptop). In an informal pilot study, we implemented dynamic binaural rendering of 12th-order SMA signals on a standard laptop (Apple MacBook Pro 15 Mid 2018, Intel Core i7 2,6 GHz) using the SMATBIN filter approach in combination with the SoundScape Renderer for fast convolution and Cockos REAPER for audio playback of the multi-channel stream, resulting in a CPU load of only about 26% on average. Thus, using SMATBIN filters should also enable real-time rendering of SMA signals that are first spatially upsampled to improve the quality [23], which significantly increases the spatial order and thus the number of audio channels. However, a direct objective comparison of the SMATBIN filter approach with other real-time rendering chains such as ReTiSAR [10] or SPARTA [8] in terms of required computational power is not easily possible, as the implementations as well as the frameworks in which the renderers run differ too much to obtain meaningful results.

Overall, the SMATBIN filter method is particularly well suited for real-time binaural rendering of SMA signals in VR or AR applications where the setup is fixed and the focus is on computationally efficient and pristine binaural reproduction of the sound field. Similar to recordings with a dummy head, the SMATBIN filter approach does not allow any further processing of the sound field. Thus, live concert streaming or VR teleconferencing, for example, which require no further processing and typically only a limited range of head orientations, could benefit from the presented method. For related consumer applications, which often do not require any flexible change in setup, the SMATBIN filter approach could even be embedded in a hardware system, enabling highly efficient binaural rendering of SMA signals in real-time.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13636-021-00224-5.

Additional file 1: The supplementary material provides additional results figures.

Acknowledgements

Not applicable.

Authors' contributions

All authors contributed equally to this work. JMA, TL, and CP implemented the method in Matlab and drafted the manuscript. JMA and TL created all the figures. TL and JMA implemented the Matlab and SoundScape Renderer demos. CP conceptualized the work. All authors read and approved the final manuscript.

Funding

The authors acknowledge funding by the German Federal Ministry of Education and Research (BMBF) under project reference 03FH014IX5-NarDasS and by the by European Regional Development Fund (ERDF) under project reference EFRE-0801444. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The supplementary material (Additional file 1) provides further results plots. The Matlab implementation of the proposed method is available on https://github.com/AudioGroupCologne/SMATBIN. The repository provides functions to calculate SMATBIN filters for arbitrary rigid sphere array configurations and head orientations as well as functions to generate results plots. Furthermore, the repository includes demo implementations for binaural rendering of simulated and measured SMA data using the proposed SMATBIN filter approach as well as an integration example demonstrating real-time binaural rendering of a commercially available SMA using SMATBIN filters and the SoundScape Renderer [39].

Declarations

Consent for publication

All authors agree to the publication in this journal.

Competing interests

The authors declare that they have no competing interests.

Received: 18 May 2021 Accepted: 30 September 2021
Published online: 06 November 2021

References

1. A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, B. Rafaely, Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution. J. Acoust. Soc. Am. **133**(5), 2711–2721 (2013). https://doi.org/10.1121/1.4795780

2. B. Bernschütz, A. V. Giner, C. Pörschmann, J. M. Arend, Binaural Reproduction of Plane Waves With Reduced Modal Order. Acta Acust. united Ac. **100**(5), 972–983 (2014). https://doi.org/10.3813/AAA.918777
3. B. Bernschütz, *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording*. (Doctoral dissertation, TU Berlin, 2016). https://doi.org/10.14279/depositonce-5082
4. J. Ahrens, C. Andersson, Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre. J. Acoust. Soc. Am. **145**(4), 2783–2794 (2019). https://doi.org/10.1121/1.5096164
5. T. Lübeck, H. Helmholz, J. M. Arend, C. Pörschmann, J. Ahrens, Perceptual Evaluation of Mitigation Approaches of Impairments due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data. J. Audio Eng. Soc. **68**(6), 428–440 (2020). https://doi.org/10.17743/jaes.2020.0038
6. F. Zotter, M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording*. (Springer, Cham, Switzerland, 2019), p. 210. https://doi.org/10.1007/978-3-030-17207-7
7. H. Helmholz, C. Andersson, J. Ahrens, in *Proc. of the 45th DAGA*. Real-Time Implementation of Binaural Rendering of High-Order Spherical Microphone Array Signals (Deutsche Gesellschaft für Akustik e.V. (DEGA), Berlin, 2019), pp. 1462–1465
8. L. McCormack, A. Politis, in *Proc. of the AES International Conference on Immersive and Interactive Audio*. SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods (Audio Engineering Society, Inc., New York, 2019), pp. 1–12
9. IEM Plugin Suite. https://plugins.iem.at. Accessed 09 Sept 2021
10. H. Helmholz, T. Lübeck, J. Ahrens, S. V. A. Garí, D. L. Alon, R. Mehra, in *Proc. of the 46th DAGA*. Updates on the Real-Time Spherical Array Renderer (ReTiSAR) (Deutsche Gesellschaft für Akustik e.V. (DEGA), Berlin, 2020), pp. 1169–1172
11. B. Rafaely, *Fundamentals of Spherical Array Processing*. (Springer, Berlin, Germany, 2015), p. 193. https://doi.org/10.1007/978-3-662-45664-4
12. A. McKeag, D. McGrath, in *Proc. of the 6th AES Australian Regional Convention*. Sound Field Format to Binaural Decoder with Head Tracking (Audio Engineering Society, Inc., New York, 1996), pp. 1–9
13. J.-M. Jot, V. Larcher, J.-M. Pernaux, in *Proc. of the 16th International AES Conference on Spatial Sound Reproduction*. A Comparative Study of 3-D Audio Encoding and Rendering Techniques (Audio Engineering Society, Inc., New York, 1999), pp. 281–300
14. M. Noisternig, A. Sontacchi, T. Musil, Robert Höl, in *Proc. of the 24th AES International Conference on Multichannel Audio - The New Reality*. A 3D Ambisonics Based Binaural Sound Reproduction System (Audio Engineering Society, Inc., New York, 2003), pp. 1–5
15. F. Zotter, M. Frank, All-round ambisonic panning and decoding. J. Audio Eng. Soc. **60**(10), 807–820 (2012)
16. C. Hold, H. Gamper, V. Pulkki, N. Raghuvanshi, I. J. Tashev, in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Improving Binaural Ambisonics Decoding by Spherical Harmonics Domain Tapering and Coloration Compensation (IEEE, New York, 2019), pp. 261–265. https://doi.org/10.1109/ICASSP.2019.8683751
17. Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, Spectral equalization in binaural signals represented by order-truncated spherical harmonics. J. Acoust. Soc. Am. **141**(6), 4087–4096 (2017)
18. mh acoustics, em32 Eigenmike. https://mhacoustics.com/products. Accessed 09 Sept 2021
19. Zylia, ZM-1. https://www.zylia.co/zylia-zm-1-microphone. Accessed 09 Sept 2021
20. O. Moschner, D. T. Dziwis, T. Lübeck, C. Pörschmann, in *Proc. of the 148th AES Convention*. Development of an Open Source Customizable High Order Rigid Sphere Microphone Array (Audio Engineering Society, Inc., New York, 2020), pp. 1–5
21. P. Stade, B. Bernschütz, M. Rühl, in *Proc. of the 27th Tonmeistertagung - VDT International Convention*. A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios (Verband Deutscher Tonmeister e.V. (VDT), Cologne, 2012), pp. 1–17
22. T. Lübeck, J. M. Arend, C. Pörschmann, in *Proc. of the 47th DAGA*. A High-Resolution Spatial Room Impulse Response Database (Deutsche Gesellschaft für Akustik e.V. (DEGA), Berlin, 2021), pp. 1–4
23. C. D. Salvador, S. Sakamoto, J. Treviño, Y. Suzuki, in *Proc. of the AES International Conference on Audio for Virtual and Augmented Reality (AVAR)*. Enhancing binaural reconstruction from rigid circular microphone array recordings by using virtual microphones (Audio Engineering Society, Inc., New York, 2018), pp. 1–9
24. Z. Ben-Hur, J. Sheaffer, B. Rafaely, Joint sampling theory and subjective investigation of plane-wave and spherical harmonics formulations for binaural reproduction. Appl. Acoust. **134**, 138–144 (2018). https://doi.org/10.1016/j.apacoust.2018.01.016
25. G. Enzner, M. Weinert, S. Abeling, J.-M. Batke, P. Jax, in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Advanced System Options for Binaural Rendering of Ambisonics Format (IEEE, New York, 2013), pp. 251–255. https://doi.org/10.1109/ICASSP.2013.6637647
26. B. Rafaely, M. Kleider, Spherical Microphone Array Beam Steering Using Wigner-D Weighting. IEEE Signal Process. Lett. **15**, 417–420 (2008). https://doi.org/10.1109/LSP.2008.922288
27. J. Ivanic, K. Ruedenberg, Rotation Matrices for Real Spherical Harmonics. Direct Determination by Recursion. J. Phys. Chem. **100**(15), 6342–6347 (1996). https://doi.org/10.1021/jp953350u
28. B. Bernschütz, C. Pörschmann, S. Spors, S. Weinzierl, in *Proc. of the International Conference on Spatial Audio (ICSA)*. SOFiA Sound Field Analysis Toolbox (Verband Deutscher Tonmeister e.V. (VDT), Cologne, 2011), pp. 8–16
29. B. Bernschütz, in *Proc. of the 39th DAGA*. A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100 (Deutsche Gesellschaft für Akustik e.V. (DEGA), Berlin, 2013), pp. 592–595
30. B. Bernschütz, C. Pörschmann, S. Spors, S. Weinzierl, in *Proc. of the 37th DAGA*. Soft-Limiting der modalen Amplitudenverstärkung bei sphärischen Mikrofonarrays im Plane Wave Decomposition Verfahren (Soft Limiting of the Modal Amplitude Gain for Spherical Microphone Arrays Using the Plane Wave Decomposition Method) (Deutsche Gesellschaft für Akustik e.V. (DEGA), Berlin, 2011), pp. 661–662
31. F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, S. Weinzierl, A Cross-Evaluated Database of Measured and Simulated HRTFs Including 3D Head Meshes, Anthropometric Features, and Headphone Impulse Responses. J. Audio Eng. Soc. **67**(9), 705–718 (2019). https://doi.org/10.17743/jaes.2019.0024
32. J. M. Arend, F. Brinkmann, C. Pörschmann, Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions. J. Audio Eng. Soc. **69**, 104–117 (2021). https://doi.org/10.17743/jaes.2020.0070
33. M. Slaney, Auditory Toolbox: A Matlab Toolbox for Auditory Modeling Work, Version 2. Technical Report #1998-010. Interval Research Corporation (1998). https://engineering.purdue.edu/~malcolm/interval/1998-010/
34. F. Brinkmann, S. Weinzierl, in *Proc. of the AES Conference on Audio for Virtual and Augmented Reality (AVAR)*. Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition (Audio Engineering Society, Inc., New York, 2018), pp. 1–10
35. B. Bernschütz, C. Pörschmann, S. Spors, S. Weinzierl, in *Proc. of the 36th DAGA*. Entwurf und Aufbau eines variablen sphärischen Mikrofonarrays für Forschungsanwendungen in Raumakustik und Virtual Audio (Design and Construction of a Variable Spherical Microphone Array for Research in Room Acoustics and Virtual Audio) (Deutsche Gesellschaft für Akustik e.V. (DEGA), Berlin, 2010), pp. 717–718
36. J. R. Driscoll, D. M. Healy, Computing Fourier Transforms and Convolutions on the 2-Sphere. Adv. Appl. Math. **15**(2), 202–250 (1994). https://doi.org/10.1006/aama.1994.1008
37. M. J. Mohlenkamp, A Fast Transform for Spherical Harmonics. J. Fourier Anal. Appl. **5**, 158–184 (1999). https://doi.org/10.1007/bf01261607
38. A. Lindau, S. Weinzierl, in *Proc. of the EAA Symposium on Auralization*. On the Spatial Resolution of Virtual Acoustic Environments for Head Movements in Horizontal, Vertical and Lateral Direction (European Acoustics Association (EAA), Espoo, 2009), pp. 1–6
39. M. Geier, J. Ahrens, S. Spors, in *Proc. of the 124th AES Convention*. The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods (Audio Engineering Society, Inc., New York, 2008), pp. 1–6
40. A. Neidhardt, F. Klein, N. Knoop, T. Köllmer, in *Proc. of the 142nd AES Convention*. Flexible Python tool for dynamic binaural synthesis applications (Audio Engineering Society, Inc., New York, 2017), pp. 1–5
41. E. Rasumow, M. Blau, S. Doclo, S. Van De Par, M. Hansen, D. Puschel, V. Mellert, Perceptual Evaluation of Individualized Binaural Reproduction Using a Virtual Artificial Head. J. Audio Eng. Soc. **65**(6), 448–459 (2017). https://doi.org/10.17743/jaes.2017.0012

42. M. Fallahi, M. Hansen, S. Doclo, S. Van De Par, D. Püschel, M. Blau, Evaluation of head-tracked binaural auralizations of speech signals generated with a virtual artificial head in anechoic and classroom environments. Acta Acust. **5**(30), 1–18 (2021). https://doi.org/10.1051/aacus/2021025
43. L. Madmoni, J. Donley, V. Tourbabin, B. Rafaely, in *Proc. of the AES International Conference on Audio for Virtual and Augmented Reality (AVAR)*. Beamforming-based Binaural Reproduction by Matching of Binaural Signals (Audio Engineering Society, Inc., New York, 2020), pp. 1–8

Publisher's Note

5.3 Perceptual Evaluation of Mitigation Approaches of Impairments due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data: Dry Acoustic Environments

Lübeck, T., Helmholz, H., Arend, J. M., Pörschmann, C., & Ahrens, J. (2020). In *Proc. of the 23rd International Conference on Digital Audio Effects (DAFx2020), Vienna, Austria* (pp. 250–257).

PERCEPTUAL EVALUATION OF MITIGATION APPROACHES OF IMPAIRMENTS DUE TO SPATIAL UNDERSAMPLING IN BINAURAL RENDERING OF SPHERICAL MICROPHONE ARRAY DATA: DRY ACOUSTIC ENVIRONMENTS

Tim Lübeck, Johannes M. Arend, Christoph Pörschmann *

Institute of Communications Engineering
TH Köln - University of Applied Sciences,
50678 Cologne, Germany
tim.luebeck@th-koeln.de

Hannes Helmholz, Jens Ahrens †

Division of Applied Acoustics
Chalmers University of Technology
412 96 Gothenburg, Sweden
hannes.helmholz@chalmers.se

ABSTRACT

Employing a finite number of discrete microphones, instead of a continuous distribution according to theory, reduces the physical accuracy of sound field representations captured by a spherical microphone array. For a binaural reproduction of the sound field, a number of approaches have been proposed in the literature to mitigate the perceptual impairment when the captured sound fields are reproduced binaurally. We recently presented a perceptual evaluation of a representative set of approaches in conjunction with reverberant acoustic environments. This paper presents a similar study but with acoustically dry environments with reverberation times of less than $0.25\,\mathrm{s}$. We examined the Magnitude Least-Squares algorithm, the Bandwidth Extraction Algorithm for Microphone Arrays, Spherical Head Filters, spherical harmonics Tapering, and Spatial Subsampling, all up to a spherical harmonics order of 7. Although dry environments violate some of the assumptions underlying some of the approaches, we can confirm the results of our previous study: Most approaches achieve an improvement whereby the magnitude of the improvement is comparable across approaches and acoustic environments.

1. INTRODUCTION

Spherical microphone arrays (SMAs) allow for capturing sound fields including spatial information. The captured sound fields can be rendered binaurally if the head-related transfer functions (HRTFs) are available on a sufficiently dense grid. Mathematically, this is performed by means of spherical harmonics (SH) expansion of the sound field and the HRTFs [1, 2]. Conceptually, it is equivalent to bringing the listener's head virtually into the sound field captured with the array. Rotation of the HRTFs relative to the sound field according to the instantaneous head orientation of the listener allows for dynamic presentation.

The physical accuracy that can be achieved with SMAs is limited, mainly due to the employment of a finite number of microphones as opposed to the continuous distribution that the theory assumes. This leads to spatial undersampling of the captured sound field, which 1) induces spatial aliasing and 2) limits the maximum obtainable SH order representation. The order of the SH presentation directly corresponds to the spatial resolution of the captured sound field. Both phenomenons can lead to audible artifacts. Another practical impairment is caused by self-noise of the microphones in the array. Studying this aspect is beyond the scope of the present paper. We refer the reader to [3, 4].

In recent years, several approaches to mitigate such impairments in binaural rendering of undersampled SMA data have been proposed. We recently conducted a listening experiment to study the perceptual effects of the mitigation approaches [5]. The study employed the acoustic data of two rooms with a reverberation time of more than $1\,\mathrm{s}$. In this contribution we present the results for a similar study, whereby the employed acoustic environments exhibit shorter reverberation times of less than $0.25\,\mathrm{s}$.

2. SPATIAL UNDERSAMPLING

To outline the phenomenon of spatial undersampling, we briefly summarize the fundamental concept of binaural rendering of SMA data. For a more detailed explanation please refer to [2, 6]. The sound pressure $S(r, \phi, \theta, \omega)$ captured by the microphones on the array surface Ω is represented in the SH domain using the spherical Fourier transform (SFT)

$$S_{nm}(r, \omega) = \int_{\Omega} S(r, \phi, \theta, \omega)\, Y_n^m(\theta, \phi)^*\, \mathrm{d}A_\Omega\,, \qquad (1)$$

whereby r denotes the array radius, ϕ and θ the azimuth and colatitude of a point on the array surface, and $\omega = 2\pi f$ the angular frequency. $Y_n^m(\theta, \phi)$ denotes the orthogonal SH basis functions for certain orders n and modes m and $(\cdot)^*$ the complex conjugate.

Based on knowledge of the sound field SH coefficients S_{nm}, the sound field on the array surface can be decomposed into a continuum of plane waves impinging from all possible directions

$$D(\phi, \theta, \omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n\, S_{nm}(r, \omega)\, Y_n^m(\phi, \theta)\,, \qquad (2)$$

with a set of radial filters d_n. Note that $S(r, \phi, \theta, \omega)$ and $D(\phi, \theta, \omega)$ do not necessarily represent the same sound fields. A SMA can incorporate a scattering body whose effect is contained in $S(r, \phi, \theta, \omega)$ but not in $D(\phi, \theta, \omega)$ where it is removed by the radial filters.

A HRTF $H(\phi, \theta, \omega)$ can be interpreted as the spatio-temporal transfer function of a plane wave to the listeners' ears. The binaural signals $B(\omega)$ for the left or right ear due to the plane wave components $D(\phi, \theta, \omega)$ impinging on the listener's head can therefore

* This work was partly supported by ERDF (European Regional Development Fund).

† This work was partly supported by Facebook Reality Labs.

be computed by weighting all HRTFs $H(\omega)$ with the plane wave coefficients of $D(\phi_d, \theta_d, \omega)$ and integrating over all propagation directions

$$B(\omega) = \frac{1}{4\pi} \int_{\Omega} H(\phi, \theta, \omega)\, D(\phi, \theta, \omega)\, \mathrm{d}A_{\Omega}\,. \tag{3}$$

Transforming the HRTFs into the SH domain as well and exploiting the orthogonality property of the SH basis functions allows to resolve the integral and compute the binaural signals for either ear as [1]

$$B(\omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} d_n\, S_{nm}(\omega, r)\, H_{nm}(\omega)\,. \tag{4}$$

The exact formulation of Eq. (4) depends on the particular definition of the employed SH basis functions [7, p. 7].

So far, we have assumed a continuously and ideally sampled sound pressure distribution on the array surface. In this case, the computation of the ear signals is perfect i.e., $B(\omega)$ in (4) are the signals that arise if the listener with HRTFs $H(\phi, \theta, \omega)$ is exposed to the sound field that the microphone array captures. Real-world SMAs employ only a finite number of discrete microphones. As a result, spatial aliasing and truncation of the SH order n occur, which makes the ear signals that are computed by the processing pipeline differ from the true ones. This can significantly affect the perceptual quality of binaural reproduction, as shown by numerous research [2, 8, 9, 10]. These impairments due to spatial undersampling are briefly discussed in the following.

2.1. Spatial Aliasing

Similar to time-frequency sampling, where frequency components above the Nyquist-frequency are aliased to lower frequency regions, sampling the space with a limited number of sensors introduces spatial aliasing. Note that this applies for both, sampling of the sound field $S(\cdot)$ as well as for the sampling of the HRTFs $H(\cdot)$. In case aliasing occurs, higher spatial modes cannot be reliably resolved and leak into lower modes. Generally, higher modes are required for resolving high frequency components with smaller wavelengths. Spatial aliasing therefore limits the upper bound of the time-frequency bandwidth that can be deduced reliably from the array signals. While theoretically being apparent at all temporal frequencies f, spatial aliasing artifacts are considerable only above the temporal-frequency [6]

$$f_{\mathrm{A}} = \frac{N_{\mathrm{sg}}\, c}{2\pi r}\,. \tag{5}$$

Thereby, c denotes the speed of sound and N_{sg} the maximum resolvable SH order n of the sampling scheme. The leakage of higher spatial modes into lower spatial modes results in an increase of the magnitudes at temporal-frequencies above f_{A}. Although spatial aliasing primarily impairs spatial properties, it therefore also affects the time-frequency spectrum of the binaural signals.

2.2. Spherical Harmonic Truncation

Orthogonality of the SH basis functions $Y_n^m(\cdot)$ is given only up to the order $n = N_{\mathrm{sg}}$ (Eq. (5)) due to the discrete sampling of the SMA surface. Spatial modes for $n > N_{\mathrm{sg}}$ are spatially distorted and are ordinarily not computed. This order truncation results in a loss of spatial information. The sampling of the SMA is usually sparser than that of the HRTFs so that the SMA is the limiting factor.

Also the spatial order truncation affects the time-frequency representation by discarding components with mostly high frequency content. In addition, hard truncation of the SH coefficients at a certain order n results in side-lobes in the plane wave spectrum in Eq. (2) [11], which can further impair the binaural signals.

3. MITIGATION APPROACHES

In the last years, a number of different approaches to improve binaural rendering of SMA captures have been presented in the literature. In the following, a selection of approaches is summarized. These are the approaches that we evaluated in the experiment presented in Sec. 6.

3.1. Pre-Processing of Head-Related Transfer Functions

Since in practice, the SH order truncation of high-resolution HRTFs cannot be avoided, a promising approach to mitigate the truncation artifacts is to pre-process the HRTFs in such a way that the major energy is shifted to lower orders without notably decreasing the perceptual quality. Several approaches to achieve this have been introduced. A summary of a selection of pre-processing techniques is presented in [12]. In this paper, we investigate two concepts.

3.1.1. Spatial Subsampling

For the spatial subsampling method [2] (SubS), the HRTFs are transformed into the SH domain up to the highest SH order N_{sg} that the sampling grid supports. Based on this representation, the HRTFs are spatially resampled with a reduced maximum SH order N'_{sg} to the grid on which the sound field is sampled, which is usually more coarse.

This process modifies the spatial aliasing in the signals in a favorable way [2]. Fig. 1 depicts the energy distribution of dummy head HRTFs [13] with respect to SH order (y-axis) and frequency (x-axis). The left-hand diagram illustrates the untreated HRTFs with a significant portion of energy at high SH orders. The middle diagram shows the same HRTF set being subsampled to a 5th-order Lebedev grid. Evidently, the information can be reliably obtained only up to the 5th order.

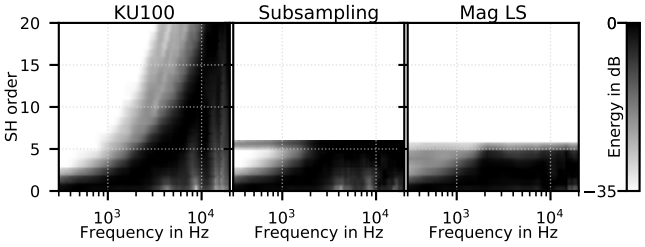


Figure 1: Energy distribution in dB with respect to order and frequency of the HRTFs of a Neumann KU100 dummy head. Untreated (left), subsampled (center), MagLS pre-processed (right).

3.1.2. Magnitude Least-Squares

Another HRTF pre-processing approach is the Magnitude Least-Squares (MagLS) [14] algorithm, which is an improvement of the Time Alignment (TA) proposed by the same authors. Both approaches are based on the duplex theory [15]. At high frequencies, the interaural level differences (ILDs) become perceptually more relevant than the interaural time differences (ITDs). However, at high frequencies, the less relevant phase information constitutes a major part of the energy. Thus, removing the linear phase at high frequencies decreases the energy in high modes, without losing relevant perceptual information. MagLS aims to find an optimum phase by solving a least-squares problem that minimizes the differences in magnitude to a reference HRTF set, resulting in minimal phase in favor of optimal ILDs. Fig. 1 (right) illustrates the energy distribution of MagLS pre-processed HRTFs for SH order 5. The major part of the energy is shifted to SH coefficients of orders below 5.

The major difference between both HRTF pre-processing approaches is that subsampling results in a HRTF set defined for a reduced number of directions and thus allowing only for a limited SH representation. In contrast, MagLS does not change the HRTF sampling grid and thus, theoretically, allows expansion up to the original SH order.

3.2. Bandwidth Extension Algorithm for Microphone Arrays

Besides pre-processing of the HRTFs, there are algorithms that are applied to the sound field SH coefficients. The Bandwidth Extension Algorithm for Microphone Arrays (BEMA) [16, 2] synthesizes the SH coefficients at $f \geq f_\mathrm{A}$ by extracting spatial and spectral information from components $f < f_\mathrm{A}$. The time-frequency spectral information is obtained by an additional omnidirectional microphone in the center of the microphone array (which is evidently not feasible in practice if a scattering object is employed). The BEMA coefficients can then be estimated as the combination of spatial and spectral information.

Fig. 2 depicts the magnitudes of plane wave components calculated for a broadband plane wave impinging from $\phi = 180°$, $\theta = 90°$ on a 50 sampling point Lebedev grid SMA with respect to azimuth angle (x-axis) and frequency (y-axis). The top diagram is based on untreated SH coefficients, the bottom diagram illustrates the effect of BEMA. For the example of a single plane wave, the sound field is perfectly reconstructed over the entire audible bandwidth.

3.3. Spherical Harmonic Tapering

SH order truncation induces side-lobes in the plane wave spectrum, which can be reduced by tapering high orders n [11]. In other words, an order-dependent scaling factor is applied to all SH modes and coefficients of that order. Different windows have been discussed, and a cosine-shaped fade-out was found to be the optimal choice. Additionally, the authors recommend to equalize the binaural signals with the so-called Spherical Head Filter, as discussed in the subsequent section. The combination of SH tapering and spherical head filters is referred to as Tap+SHF in the remainder.

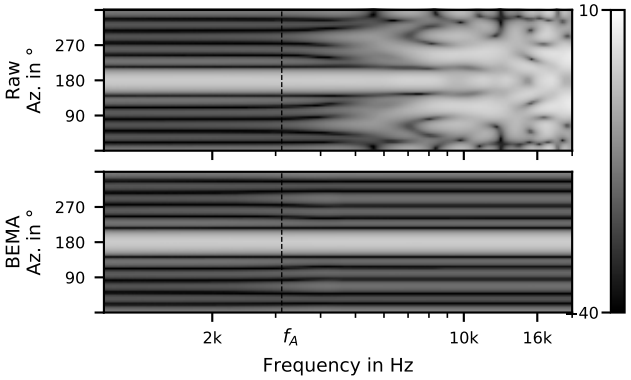


Figure 2: Plane wave magnitudes of a plane wave impact from $\phi = 180°$, $\theta = 90°$ on a 50 sampling point Lebedev grid SMA with a radius of $8.75\,\mathrm{cm}$. The top diagram depicts the untreated magnitudes, the bottom diagram the plane wave calculated after BEMA processing.

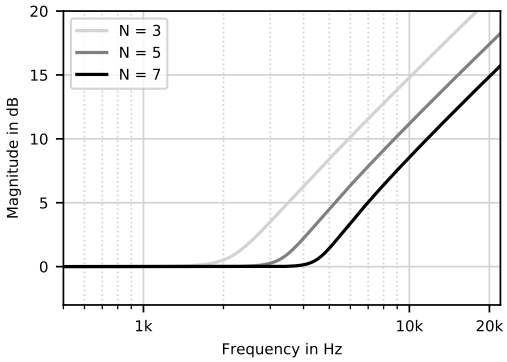


Figure 3: Spherical Head Filter (SHF) for orders $N = (3, 5, 7)$.

3.4. Spectral Equalization

The modification of the time-frequency response due to spatial undersampling is a perceptually distinctive impairment, as shown e.g. in [10]. Therefore, a third category of mitigation approaches is global equalization of the binaural signals. Different approaches have been introduced in the literature to design such equalization filters. The Spherical Head Filter (SHF) [8] compensates for the low-pass behavior of SH order truncation. The authors disregard spatial aliasing effects and proposed a filter based on the plane wave density function of a diffuse sound field. The resulting filters for different SH orders are depicted in Fig. 3. A similar approach to equalize this low-pass effect has been discussed in [17]. In the following we investigate the SHFs.

4. EMPLOYED DATA

The stimuli in our study were created from measured array room impulse responses using the `sound_field_analysis-py` Python toolbox [18] and the impulse response data set from [19]. This data set contains both binaural room impulse responses (BRIRs) measured with a Neumann KU100 dummy head as well as array room impulse responses (ARIRs) captured on various Lebedev grids under identical conditions. This allows for a direct

Figure 4: The Control Room 1 (left) and 7 (right) (CR1, CR7) with reverberation times of less than $0.25\,\mathrm{s}$ (measured at $500\,\mathrm{Hz}$ and $1\,\mathrm{kHz}$) at the WDR Broadcast studios, that were auralized in the listening experiment.

comparison of binaural auralization of SMA data to the ground truth dummy head data. The ARIR measurements were performed with the VariSphear device [20], which is a fully automated robotic measurement system that sequentially captures directional impulse responses on a spherical grid for emulating a SMA. To obtain impulse responses of a rigid sphere array, the Earthworks M30 microphone was flush-mounted in a wooden spherical scattering body (see [19, Fig. 12]). All measurements were performed in four different rooms at the WDR broadcast studios in Cologne, Germany. In this study we employ the measurement data of the rooms Control Room 1 (CR1) and Control Room 7 (CR7) (Fig. 4), which both have short reverberation times of less than $0.25\,\mathrm{s}$. Recall that we conducted a similar study with the rooms Small Broadcast Studio (SBS) and Large Broadcast Studio (LBS) with approximate reverberation times of $1\,\mathrm{s}$ and $1.8\,\mathrm{s}$ in [5].

The Neumann KU100 HRIR set, measured on a 2702 sampling point Lebedev grid [13], is used to synthesize binaural signals $B(\omega)$ for a pure horizontal grid of head orientations with 1° resolution based on ARIRs according to Eq. (4). We denote this data "ARIR renderings" in the following. Likewise, the BRIRs of the dummy head are available for the same head orientations so that a direct comparison of both auralizations is possible.

In order to restrict the gain of the radial filters $d_n(\omega)$ in (4), we employ a soft-limiting approach [2, pp. 90-118]. Fig. 5 illustrates the influence of the soft-limiting for the left-ear binaural room transfer functions (BRTFs) resulting from a broadband plane wave impinging from $(\phi = 0^\circ, \theta = 90^\circ)$ on a simulated 2702 sampling point Lebedev SMA. The BRTFs were calculated up to the 35th-order using the different radial filter limits 0, 10, 20, and $40\,\mathrm{dB}$. It can be seen that a limit of $0\,\mathrm{dB}$ leads to a significant attenuation of the high frequency components, but provides an advantageous signal-to-noise ratio in the resulting ear signals nevertheless [2, 4]. Although this is not required for the ideal rendering conditions in this study, we chose $0\,\mathrm{dB}$ soft-limiting for this contribution in order to produce comparable results to previous studies [2, 10].

All mitigation algorithms were implemented with `sound_field_analysis-py` [18]. Solely the MagLS HRIRs were pre-processed with MATLAB code provided by the authors of [14]. Every ARIR parameter set was processed with each of the mitigation algorithms MagLS, Tapering+SHF, SHF, and SubS (Spatial Subsampling), as well as an untreated (Raw) ARIR rendering was produced.

Previous studies showed that SH representations of an order of less than 8 exhibit audible undersampling artifacts, i.e., a clear perceptual difference to the reference dummy head data [10]. Since

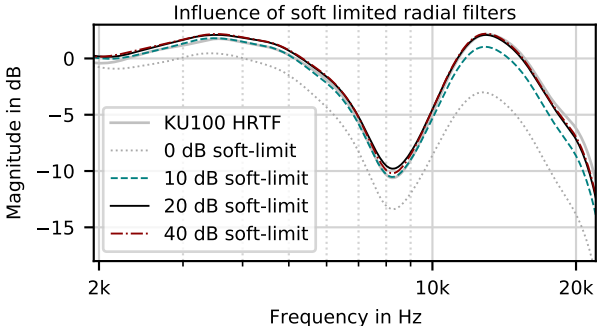


Figure 5: Left ear magnitude responses of the frontal KU100 HRTF, and ARIR binaural renderings up to order 35 involving radial filters with different soft-limits. The ARIR renderings are based on a simulated broadband plane wave impinging virtual 2702 Lebedev SMAs from $(\phi = 0^\circ, \theta = 90^\circ)$. The deviation to the magnitudes of the HRTF illustrates the influence of the soft limit. All magnitude responses are 1/3-octave-smoothed.

this work investigates the effectiveness of mitigation approaches for undersampled sound fields, we chose to focus on SH orders below 8 for the subsequent instrumental and perceptual evaluation. Significant beneficial effects of the mitigation approaches for higher orders are not expected.

5. INSTRUMENTAL EVALUATION

In this section, we compare the mitigation approaches based on 3rd SH order array data of CR7, which has a reverberation time of about $0.25\,\mathrm{s}$. We used ARIRs from a 50-point Lebedev grid. We calculated the BRIRs for 360 azimuth directions in the horizontal plane in steps of 1° and compare them to the measured ground truth dummy head BRIRs for the same head orientations.

Absolute spectral differences between dummy head and array BRIRs in dB are illustrated in Fig. 6. The top diagram depicts the deviations averaged over all 360 directions with respect to frequency (x-axis). The bottom diagram shows the differences averaged over 40 directions contralateral to the source position. It is evident that the spectral differences tend to be larger on this contralateral side.

The untreated (Raw) rendering indicated by the dashed line is clearly affected by undersampling artifacts above f_A. Around the

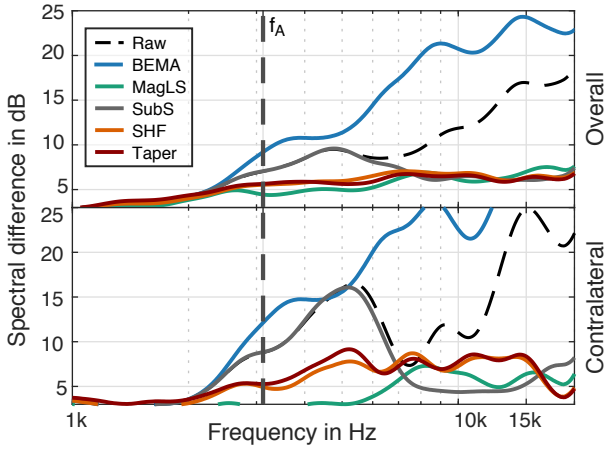


Figure 6: Absolute spectral differences of dummy head and SMA binaural signals in dB. Top: averaged over 360 horizontal directions. Bottom: averaged over 40 directions around the contralateral side.

contralateral side, these differences increase rapidly. Both HRTF pre-processing algorithms (SubS (gray) and MagLS (green)) significantly decrease the difference to the reference whereby MagLS tends to produce the lowest deviations.

Although BEMA (blue) was shown to be effective for very simple sound fields like a single plane wave, it produces significantly larger deviations from the reference than Raw. As noted by the authors of BEMA [2], even for a simple sound field composed of three plane waves from different directions and arbitrary phase, BEMA introduces audible comb filtering artifacts. Additionally, the averaging of the SH coefficients from lower modes to extract the spatial information for higher modes, leads to a perceivable low-pass effect, which produces the large differences towards higher frequencies.

The SHFs and Tapering perform comparably. Both methods employ global filtering to the binaural signals. The differences at the contralateral side are larger than for frontal directions.

6. PERCEPTUAL EVALUATION

Some of the approaches considered here have already been perceptually evaluated in listening experiments. Subsampling showed to significantly improve the perceptual quality [2], although it provokes stronger spatial aliasing. Time Alignment, Subsampling and SHFs were compared in [9]. The results showed that mostly Time Alignment, which is a predecessor of MagLS, yields better results than Subsampling. The SHFs were rated worst of the three tested methods, matching the instrumental results depicted in Fig. 6. This may be due to the fact that global equalization shifts the error in binaural time-frequency spectra to lateral directions. The perceptual evaluation of BEMA showed improvements when auralizing simulated sound fields with a limited number of sound sources [2]. However, for measured diffuse sound fields, BEMA introduces significant artifacts and thus is no promising algorithm for real-world applications. To our knowledge, Tapering has not been evaluated perceptually in a formal manner.

6.1. Methods

6.1.1. Stimuli

The stimuli were calculated as described in Sec. 4 for the SH orders 3, 5 and 7 for 360 directions along the horizontal plane with steps of 1° for the room CR7 and CR1. The 3rd and 5th-order renderings are based on impulse response measurements on the 50 sampling point Lebedev grid while for order 7 the 86 sampling point Lebedev grid was used. Previous studies showed strong perceptual differences between ARIR and dummy head auralizations in particular for lateral sound sources [9, 10]. Therefore, each ARIR rendering was generated for a virtual source in the front ($\phi = 0°, \theta = 90°$) and at the side ($\phi = 90°, \theta = 90°$). To support transparency, static stimuli for both tested sound source positions are publicly available [1]. Anechoic drum recordings were used as the test signal in particular because drums have a wide spectrum and strong transients making them a critical test signal. Previous studies showed that certain aspects are only induced with critical signals [2, 10].

6.1.2. Setup

The experiment was conducted in a quiet acoustically damped audio laboratory at Chalmers University of Technology. The SoundScape Renderer (SSR) [21] in binaural room synthesis (BRS) mode was used for dynamic auralization. It convolves arbitrary input test signals with a pair of BRIRs corresponding to the instantaneous head orientation of the listener, which was tracked along the azimuth with a Polhemus Patriot tracker. The binaural renderings were presented to the participants using AKG K702 headphones with a Lake People G109 headphone amplifier at a playback level of about 66 dBA. The output signals of the SSR were routed to an Antelope Audio Orion 32 DA converter at 48 kHz sampling frequency and a buffer length of 512 samples. Equalization according to [19] was applied to the headphones and the dummy head. The entire rendering and performance of the listening experiment were done on an iMac Pro 1.1.

6.1.3. Paradigm and Procedure

The test design was based on the Multiple Stimulus with Hidden Reference and Anchor (MUSHRA) methodology proposed by the International Telecommunication Union (ITU) [22]. The participants were asked to compare the ARIR renderings to the dummy head reference in terms of overall perceived difference. The anchor consists of diotic non-head-tracked BRIRs, low-pass filtered at a cutoff at 3 kHz. Each trial, i.e., a MUSHRA page, comprised 8 stimuli to be rated by the subjects (BEMA, MagLS, SHF, Tapering+SHF, SubS, Raw, hidden reference (Ref), Anchor). The experiment was composed of 12 trials: 3 SH orders $(3, 5, 7) \times 2$ nominal source positions $(0°, 90°) \times 2$ rooms (CR1, CR7).

The subjects were provided a graphical user interface (GUI) with continuous sliders ranging from '*No difference*', '*Small difference*', '*Moderate difference*', '*Significant difference*' to '*Huge difference*' as depicted in Fig. 7.

14 participants in the age between 21 and 50 years took part in the experiment. Most of them were MSc students or staff at the Division of Applied Acoustics of Chalmers University of Technology. The subjects were sitting in front of a computer screen with a keyboard and a mouse. The drum signal was playing continuously,

[1] http://doi.org/10.5281/zenodo.3931629

Figure 7: Employed graphical user interface of the listening experiment.

and it was possible to listen to each stimulus as often and long as desired. The participants were allowed and strongly encouraged to move their heads during the presentation of the stimuli. At the beginning of each experiment, the subjects rated four training stimuli that covered the entire range of perceptual differences of the presented stimuli in the main part of the experiment. These training stimuli consisted of a BEMA and MagLS rendering of CR1 data at order 3 for the lateral sound source position as well as the corresponding anchor and reference. The experiment took on average about 30 minutes per participant.

6.2. Results

As recommended by the ITU [22], we post-screened all reference and anchor ratings. Two participants rated the anchor higher than 30 (44, 36). We found no further inconsistencies so that we chose not to exclude these participants.

In the listening experiment, we solely presented one order and one direction per trial. We want to therefore highlight that the direct comparison of the ratings for different orders and different source positions as well as subsequent interpretation has to be performed with reservation. All stimuli were presented in randomized order and the corresponding references and anchors were always the same for each condition so that some amount of consistency in the subject's responses may be assumed. We therefore present a statistical analysis in the following that includes comparisons between orders and positions as it is commonly performed with MUSHRA data.

Fig. 8 presents the interindividual ratings in form of boxplots. The plots are divided for each room and sound source position and present the ratings with respect to the algorithm (x-axis) and order as indicated by the color. Two major observations can be made: 1) Considering the ratings of the Raw conditions shows that mostly higher-order renderings were perceived closer to the reference than lower-order renderings. 2) The algorithms MagLS, Tapering+SHF, SubS, and SHF all improve ARIR renderings compared to untreated renderings. This improvements seems to become weaker with increasing order.

For statistical analysis of the results, a repeated measures ANOVA was performed. We applied a Lilliefors test for normality to test the assumptions for the ANOVA. It failed to reject the null hypothesis in 4 of 72 conditions at a significance level of $p = 0.05$. However, parametric tests such as the ANOVA are generally robust to violations of normality assumption [25]. For further analysis Greenhouse-Geisser corrected p-values are considered, with the associated ϵ-values for correction of the degrees of freedom of the F-distribution being reported.

A four-way repeated measures ANOVA with the within-subject factors algorithm (BEMA, MagLS, Tapering+SHF, SHF, SubS, and Raw), order $(3,\ 5,\ 7)$, room (CR1, CR7), and nominal source position $(0°,\ 90°)$ was performed. The associated mean values with respect to algorithm (x-axis), and SH order (color) are depicted in Fig. 9. Each value was calculated as the mean value of the ratings of all participants for both directions and both rooms. The $95\,\%$ within-subject confidence intervals were determined as proposed by [23, 24] based on the main effect of algorithm. Similar to the boxplots, the mean values indicate that all algorithms except BEMA yield considerable improvements.

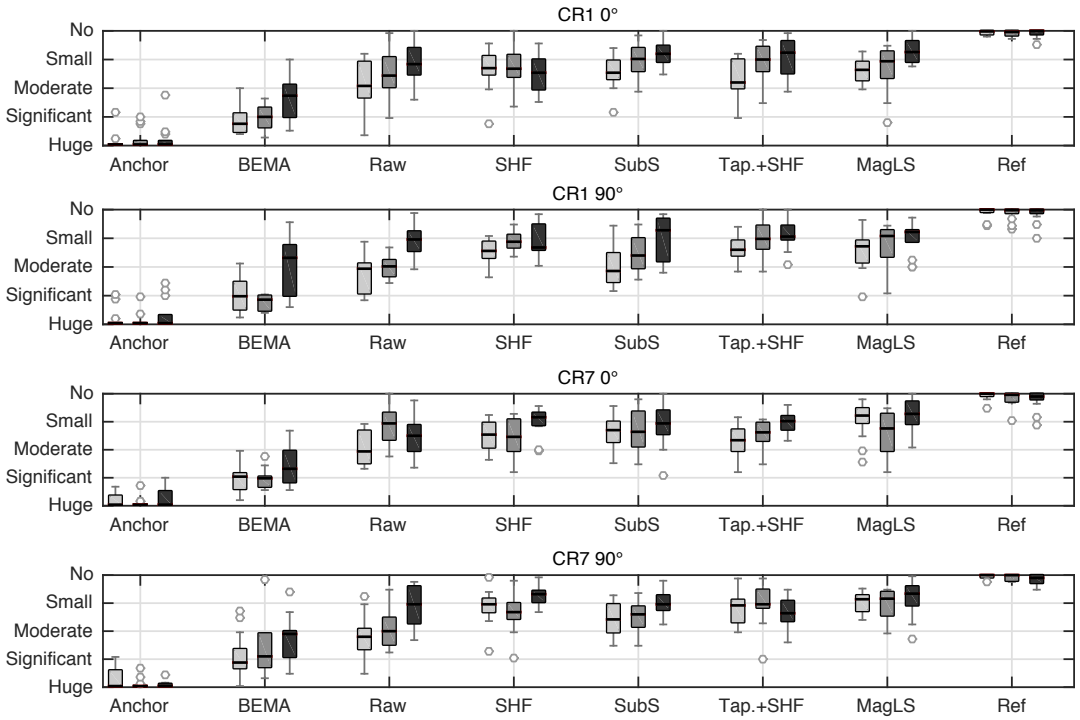
The ANOVA revealed the significant main effects algorithm ($F(5, 65) = 143.64,\ p < .001,\ \eta_p^2 = .917,\ \epsilon = .457$), and order ($F(2, 26) = 37.382, p < .001, \eta_p^2 = .742, \epsilon = .773$). These significant effects match the observations made so far. Mostly, higher-order renderings yielded smaller perceptual differences than lower-order ones. Further, the algorithm significantly influences the perceptual character of ARIR renderings. The ANOVA revealed the significant interaction of algorithm×order ($F(10, 130) = 4.756,\ p < .001,\ \eta_p^2 = .268,\ \epsilon = .556$). Thus, the algorithms seem to perform differently with respect to the rendering order. The significant effect of the interaction of algorithm×source position ($F(5, 65) = 7.176, p < .001$, $\eta_p^2 = .356, \epsilon = .774$) shows that the performance of the algorithm also depends on the sound source position.

The ANOVA also revealed two significant interactions involving the factor room: The interaction of algorithm×room ($F(5, 65) = 2.864,\ p < .040,\ \eta_p^2 = .181,\ \epsilon = .695$), as well as order×room ($F(2, 26) = 4.736, p < .024, \eta_p^2 = .267, \epsilon = .853$) were found to be significant. The results of the listening experiment and the ANOVA values, are available as well [1].

7. DISCUSSION AND CONCLUSIONS

We presented a perceptual evaluation of approaches for mitigating the perceptual impairment due to spatial aliasing and order truncation in binaural rendering of spherical microphone array data. The present results employing dry acoustic environments together with previous results on reverberant environments [5] suggest the following:

- Bandwidth Extension Algorithm for Microphone Arrays (BEMA) is the only method that causes larger perceptual differences to the ground truth signal than without mitigation.
- Depending on the condition, all other mitigation approaches produce either no improvement or an improvement that is comparable in magnitude.
- Mitigation is more effective at lower orders and is hardly detectable at order 7.
- We did not find a dependency on the room although some mitigation approaches are based on a diffuse field assumption, which fulfilled better in more reverberant rooms.
- In both experiments Tapering+SHF was sometimes rated closer to the reference when rendered at order 5, instead of order 7. This might be caused by the cosine-shaped windowing of the Tapering algorithm, which modifies higher rendering orders more than lower ones.



b

Figure 8: Interindividual variation in the ratings of perceptual difference between the stimulus and the dummy head reference with respect to the algorithm (x-axis), and SH order (color) for each room and virtual source position separately. Each box indicates the 25th and 75th percentiles, the median value (black line), the outliers (grey circles) and the minimum / maximum ratings not identified as outliers (black whiskers).

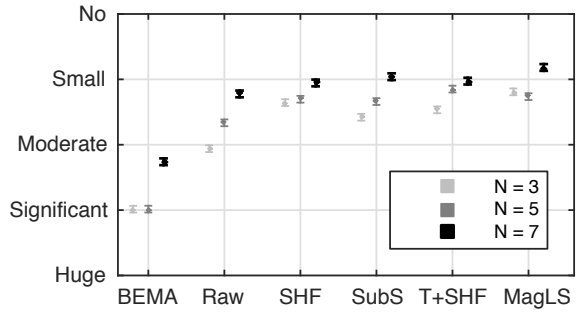


Figure 9: Mean values of the ratings pooled over both rooms with respect to the algorithm. The 95 % within-subject confidence intervals were calculated according to [23, 24]. The ratings for different SH orders are displayed separately as indicated by the color.

8. ACKNOWLEDGMENTS

We thank Christian Schörkhuber, Markus Zaunschirm, and Franz Zotter of IEM at the University of Music and Performing Arts in Graz for providing us with their code of MagLS, and all participants of the listening experiment for their support.

9. REFERENCES

[1] Boaz Rafaely and Amir Avni, "Interaural cross correlation in a sound field represented by spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 823–828, 2010.

[2] Benjamin Bernschütz, *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording*, Ph.D. thesis, Technische Universität Berlin, 2016.

[3] Hannes Helmholz, Jens Ahrens, David Lou Alon, Sebastià V. Amengual Garí, and Ravish Mehra, "Evaluation of Sensor Self-Noise In Binaural Rendering of Spherical Microphone Array Signals," in *Proc. of the IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 161–165, IEEE.

[4] Hannes Helmholz, David Lou Alon, Sebastià V. Amengual Garí, and Jens Ahrens, "Instrumental Evaluation of Sensor Self-Noise in Binaural Rendering of Spherical Microphone Array Signals," in *Forum Acusticum*, Lyon, France, 2020, pp. 1–8, EAA.

[5] Tim Lübeck, Hannes Helmholz, Johannes M. Arend, Christoph Pörschmann, and Jens Ahrens, "Perceptual Evaluation of Mitigation Approaches of Impairments due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data," *Journal of the Audio Engineering Society*, pp. 1–12, 2020.

[6] Boaz Rafaely, *Springer Topics in Signal Processing Springer Topics in Signal Processing*, Springer, 2015.

[7] Carl Andersson, "Headphone Auralization of Acoustic Spaces Recorded with Spherical Microphone Arrays," M.S. thesis, Chalmers University of Technology, 2017.

[8] Zamir Ben-Hur, Fabian Brinkmann, Jonathan Sheaffer, Stefan Weinzierl, and Boaz Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4087–4096, 2017.

[9] Markus Zaunschirm, Christian Schörkhuber, and Robert Höldrich, "Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3616–3627, 2018.

[10] Jens Ahrens and Carl Andersson, "Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre," *The Journal of the Acoustical Society of America*, vol. 145, no. April, pp. 2783–2794, 2019.

[11] Christoph Hold, Hannes Gamper, Ville Pulkki, Nikunj Raghuvanshi, and Ivan J. Tashev, "Improving Binaural Ambisonics Decoding by Spherical Harmonics Domain Tapering and Coloration Compensation," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 261–265.

[12] Fabian Brinkmann and Stefan Weinzierl, "Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition," in *Proceedings of the AES Conference on Audio for Virtual and Augmented Reality*, Redmond, USA, 2018, pp. 1–10.

[13] Benjamin Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," in *Proceedings of the 39th DAGA*, Meran, Italy, 2013, pp. 592–595.

[14] Christian Schörkhuber, Markus Zaunschirm, and Robert Holdrich, "Binaural rendering of Ambisonic signals via magnitude least squares," in *Proceedings of 44th DAGA*, Munich, Germany, 2018, pp. 339–342.

[15] Lord Rayleigh, "XII. On our perception of sound direction," *Philosophical Magazine Series 6*, vol. 13, no. 74, pp. 214–232, 1907.

[16] Benjamin Bernschütz, "Bandwidth Extension for Microphone Arrays," in *Proceedings of the 133th AES Convention*, San Francisco, USA, 2012, pp. 1–10.

[17] Thomas McKenzie, Damian T. Murphy, and Gavin Kearney, "Diffuse-Field Equalisation of binaural ambisonic rendering," *Applied Sciences*, vol. 8, no. 10, 2018.

[18] Christoph Hohnerlein and Jens Ahrens, "Spherical Microphone Array Processing in Python with the sound field analysis-py Toolbox," in *Proceedings of the 43rd DAGA*, Kiel, Germany, 2017, pp. 1033–1036.

[19] Philipp Stade, Benjamin Bernschütz, and Maximilian Rühl, "A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios," in *Proceedings of the 27th Tonmeistertagung - VDT International Convention*, Cologne, Germany, 2012, pp. 551–567.

[20] Benjamin Bernschütz, Christoph Pörschmann, Sascha Spors, and Stefan Weinzierl, "Entwurf und Aufbau eines variablen sphärischen Mikrofonarrays für Forschungsanwendungen in Raumakustik und Virtual Audio," in *Proceedings of 36th DAGA*, Berlin, Germany, 2010, pp. 717–718.

[21] Matthias Geier, Jens Ahrens, and Sascha Spors, "The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods," in *Proceedings of the 124th AES Convention*, Amsterdam, Netherlands, 2008, pp. 179–184, Code publicly available at "http://spatialaudio.net/ssr/".

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

[23] Geoffrey R Loftus, "Using confidence intervals in within-subject designs," *Psychonomic Bulletin & Review*, vol. 1, no. 4, pp. 1–15, 1994.

[24] Jerzy Jarmasz and Justin G. Hollands, "Confidence Intervals in Repeated-Measures Designs: The Number of Observations Principle," *Canadian Journal of Experimental Psychology*, vol. 63, no. 2, pp. 124–138, 2009.

[25] Jürgen Bortz and Christof Schuster, *Statistik für Human- und Sozialwissenschaftler*, Springer-Verlag, Gießen, Germany, 7 edition, 2010.

5.4 Perceptual Evaluation of Mitigation Approaches of Impairments due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data

Lübeck, T., Helmholz, H., Arend, J. M., Pörschmann, C., & Ahrens, J. (2020). *J. Audio Eng. Soc.*, *68*(6), 428–440. https://doi.org/10.17743/jaes.2020.0038

OPEN ACCESS Freely available online

T. Lübeck, H. Helmholz, J. M. Arend, C. Pörschmann, and J. Ahrens "Perceptual Evaluation of Mitigation Approaches of Impairments Due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data"
J. Audio Eng. Soc., vol. 68, no. 6, pp. 428–440, (2020 June.).
DOI: https://doi.org/10.17743/jaes.2020.0038

Perceptual Evaluation of Mitigation Approaches of Impairments Due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data

TIM LÜBECK,[1,2] *AES Student Member*, (tim.luebeck@th-koeln.de)
HANNES HELMHOLZ,[3] *AES Student Member*, (hannes.helmholz@chalmers.se)
JOHANNES M. AREND,[1,2] *AES Student Member*, (johannes.arend@th-koeln.de)
CHRISTOPH PÖRSCHMANN,[1] *AES Associate Member*, (christoph.poerschmann@th-koeln.de)
AND
JENS AHRENS,[3] *AES Member* (jens.ahrens@chalmers.se)

[1] *TH Köln – University of Applied Sciences, Cologne, Germany*
[2] *Technical University of Berlin, Berlin, Germany*
[3] *Chalmers University of Technology, Gothenburg, Sweden*

Spherical microphone arrays (SMAs) are widely used to capture spatial sound fields that can then be rendered in various ways as a virtual acoustic environment (VAE) including headphone-based binaural synthesis. Several practical limitations have a significant impact on the fidelity of the rendered VAE. The finite number of microphones of SMAs leads to spatial undersampling of the captured sound field, which, on the one hand, induces spatial aliasing artifacts and, on the other hand, limits the order of the spherical harmonics (SH) representation. Several approaches have been presented in the literature that aim to mitigate the perceptual impairments due to these limitations. In this article, we present a listening experiment evaluating the perceptual improvements of binaural rendering of undersampled SMA data that can be achieved using state-of-the-art mitigation approaches. In particular, we examined the Magnitude Least-Squares algorithm, the Bandwidth Extraction Algorithm for Microphone Arrays, Spherical Head Filters, SH Tapering, and a newly proposed equalization filter. In the experiment, subjects rated the perceived differences between a dummy head and the corresponding SMA auralization. We found that most mitigation approaches lead to significant perceptual improvements, even though audible differences to the reference remain.

0 INTRODUCTION

The increasing number of virtual and augmented reality applications creates the demand for high-fidelity virtual acoustic environments (VAEs). These can be created based on either simulations or captured data. A common method for capturing and auralizing spatial sound fields is the measurement of impulse responses with a dummy head. Such impulse responses represent the acoustic path from the sound source to the ears of a listener and are referred to as either head-related impulse responses (HRIRs), when representing anechoic conditions, or binaural room impulse responses (BRIRs), when representing nonanechoic conditions. Interactive VAEs that adapt to the listener's head orientation can be realized with head tracking based on sequential dummy head measurements on adequately high-resolution grids of head orientations.

However, this technique of sound field capture makes it impossible to realize auralizations of dynamic scenarios such as music concerts. An alternative to the time-consuming sequential dummy head measurements is a continuous capture of the sound field, including all dynamic changes. By means of a distribution of sensors in the region of interest such as a spherical microphone array (SMA), the original sound field can be reconstructed.

VAEs can be rendered to a listener with different loudspeaker-based reproduction methods such as Ambisonics [1] or wave-field synthesis [2]. In this paper, we fo-

cus on headphone-based binaural reproduction. Binaural reproduction computes the signals that would arise at the listener's ears if he/she were exposed to the sound field that the microphone array captured. This is performed by virtually exposing the listener's head to the sound field that impinges on the SMA. The method utilizes a spherical harmonics (SH) representation of the sound field as well as of a set of HRIRs (see, e.g., [3, 4]).

The physical accuracy that can be achieved with SMAs is limited, mainly due to the employment of a finite number of microphones as opposed to the continuous distribution that the theory assumes. This leads to spatial undersampling of the captured sound field, which induces spatial aliasing and limits the order of the SH representation of the captured sound field that can be obtained. The order of the SH presentation directly corresponds to the spatial resolution of the sound field. Both phenomena can lead to audible impairments.

Previous studies such as [3, 4] compared binaural auralizations based on SMA data to a reference based on dummy head measurements of the exact same scenario. It was shown that, evidently, higher-order renderings yield more similarity to the dummy head auralizations. In direct comparison, renderings with representations below order 8 were perceived as noticeably different to the synthesis with dummy head data. Furthermore, listening experiments [4] showed that these differences are evoked mainly by high-frequency components, which are those that are primarily affected by spatial undersampling.

In recent years, several approaches to mitigate such impairments in binaural rendering of undersampled SMA data have been proposed. Although most of these approaches have been evaluated independently, up to now, no comparative listening experiment of all these methods has been made. We present a listening experiment comparing the perceptual improvements that can be achieved with the state-of-the-art undersampling mitigation approaches.

The paper is organized as follows: Sec. 1 presents the fundamentals of analyzing spatial sound fields by means of SMAs and binaural rendering. We describe the artifacts introduced by spatial undersampling and the state-of-the-art rendering approaches to mitigate these artifacts. Sec. 2 introduces the materials utilized in the comparative instrumental and perceptual evaluation in Sec. 3 and Sec. 4. The results are further discussed in Sec. 4 and completed with conclusions in Sec. 5.

1 THEORY

This section presents a conceptual overview on the binaural rendering of a sound field captured by an SMA. We refer the reader to [3, 5–7] for more detailed treatments.

1.1 Binaural Rendering of Spherical Microphone Array Data

Let $S(r, \phi, \theta, \omega)$ be the sound pressure distribution on a spherical surface Ω (for example, an SMA) with respect to the radius r, the azimuth angle ϕ ranging from 0 to 2π, the colatitude θ ranging from 0 to π, and the angular frequency $\omega = 2\pi f$, whereby f denotes the temporal frequency. Any sound pressure distribution whose mathematical representation fulfills the wave equation can be transformed into the SH domain using the spatial Fourier transform (SFT) [6]

$$S_{nm}(r, \omega) = \int_{\Omega} S(r, \phi, \theta, \omega)\, Y_n^m(\theta, \phi)^*\, \mathrm{d}A_{\Omega}\,, \tag{1}$$

where $Y_n^m(\theta, \phi)$ denote a set of SH basis functions, $(\cdot)^*$ the complex conjugate, and $\int_{\Omega}(\cdot)\,\mathrm{d}A_{\Omega} = \int_0^{2\pi}\int_0^{\pi}(\cdot)\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi$ the integration over the surface of the sphere. The SHs are orthogonal basis functions of the sphere and form a complete set of solutions of the angular component of the Helmholtz equation. Furthermore, any sound field on the spherical surface can be described as a continuum of infinitely many plane waves impinging the sphere from all possible directions. The plane wave coefficients $D(\phi_d, \theta_d, \omega)$ can be computed from the SH coefficients $S_{nm}(r, \omega)$ of the sound field on the surface of an acoustically rigid sphere as [5]

$$D(\phi_d, \theta_d, \omega) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} d_n\, S_{nm}(r, \omega)\, Y_n^m(\phi_d, \theta_d)\,. \tag{2}$$

The term d_n denotes a set of radial filters that compensate for the scattering effects on the surface of the sphere. These filters can exhibit very high amplification gains that need to be restricted in practical implementations. The influence of the radial filters has been discussed extensively, e.g., [3, pp. 90–118], [8], and [5, pp. 34–38].

A head-related transfer function (HRTF) $H(\phi, \theta, \omega)$ can be interpreted as the spatiotemporal transfer function of a given broadband plane wave to the listeners' ears. Note that we omit differentiating between left-ear and right-ear HRTFs, as well as left-ear and right-ear binaural signals in the mathematical formulations for convenience. The resulting binaural signals $Y(\omega)$ can hence be calculated by weighting all plane wave coefficients $D(\phi_d, \theta_d, \omega)$ of the sound field with the corresponding HRTF $H(\omega)$ of that direction and integrating them over all possible propagation directions as

$$Y(\omega) = \frac{1}{4\pi}\int_{\Omega} H(\phi, \theta, \omega)\, D(\phi, \theta, \omega)\, \mathrm{d}A_{\Omega}\,. \tag{3}$$

Transforming the HRTFs into the SH domain as well and exploiting the orthogonality property of the SHs allows to resolve the integral in Eq. (3) and compute the binaural signals as

$$Y(\omega) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} d_n\, S_{nm}(\omega, r)\, H_{nm}(\omega)\,. \tag{4}$$

The exact formulation of Eq. (4) depends on the particular definition of the employed SH basis functions [7, pp. 7].

1.2 Spatial Undersampling

Sec. 1.1 assumed a continuous pressure distribution on the surface of the SMA. Real-world SMAs, on the other hand, employ a discrete and finite set of sound pressure

sensors. This leads to spatial undersampling of the sound field and audible impairments in the synthesized binaural VAE. These impairments can be divided into two categories, namely spatial aliasing and SH order truncation.

1.2.1 Spatial Aliasing

When sampling continuous-time signals, components above the Nyquist frequency cannot be deduced reliably and are aliased to lower frequency components [9]. Analogously, when spatially sampling space-continuous sound fields at discrete locations, higher spatial modes cannot be deduced reliably and are mirrored into lower modes. This results in spatial ambiguities and changes in the time-frequency spectrum.

In contrast to continuous-time signals that can exhibit a limited bandwidth, sound fields are not band-limited in their modal order. Spatial aliasing is therefore apparent over the entire time-frequency spectrum. There is a temporal spatial aliasing frequency f_A

$$f_A = \frac{N_{grid}\, c}{2\pi r}\,, \tag{5}$$

above which the spatial aliasing artifacts increase rapidly [10]. Thereby, c denotes the speed of sound and N_grid the maximum resolvable SH order of the sampling scheme. In other words, spatial aliasing artifacts are very small in magnitude below f_A.

1.2.2 SH Order Truncation

The second fundamental impairment of undersampled SMA data is the truncation of the natural SH order. The integral in Eq. (1) has to be discretized to Q points, corresponding to the microphone locations Ω_q. This leads to the discrete SFT

$$S_{nm}(\omega) = \sum_{q=1}^{Q} w_q\; S(\Omega_q, \omega)\; Y_n^m(\Omega_q)^*\,. \tag{6}$$

The weights w_q ensure orthogonality of the SH basis functions. The coefficients $S_{nm}(\omega)$ can be obtained only for orders $n \leq N_\mathrm{grid}$.

1.2.3 Consequences of Spatial Undersampling

Spatial aliasing depends on the density of the SMA microphone sampling scheme, whereas order truncation solely depends on the SH order. Even though both phenomena affect similar time-frequency regions, they exhibit different and sometimes even contrary effects. The compound error of spatial aliasing and truncation was termed "sparsity error" in [11]. The authors investigated the sparsity error with a focus on binaural auralization, which is summarized in the following.

Fig. 1 illustrates the energy distribution of HRTFs in different SH modes as a function of time frequency. It can be seen that the higher SH modes contain a significant fraction of the energy at higher frequencies. Order truncation leads to a loss of spatial details in the according frequency range, which may result in an impairment of the interaural level differences (ILDs), among other things. Moreover, the hard

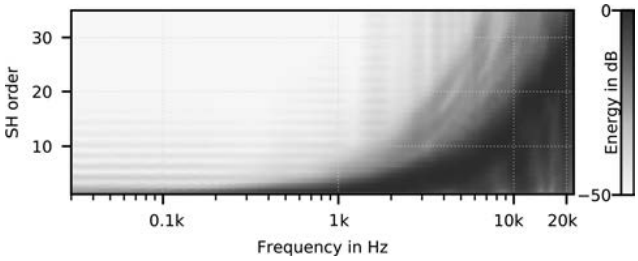


Fig. 1. Normalized logarithmic energy distribution of the head-related transfer functions (HRTFs) of the employed Neumann KU100 dummy head over frequency and SH order n. The color encodes the energy ranging from −50 dB normalized to the maximum values for each frequency bin.

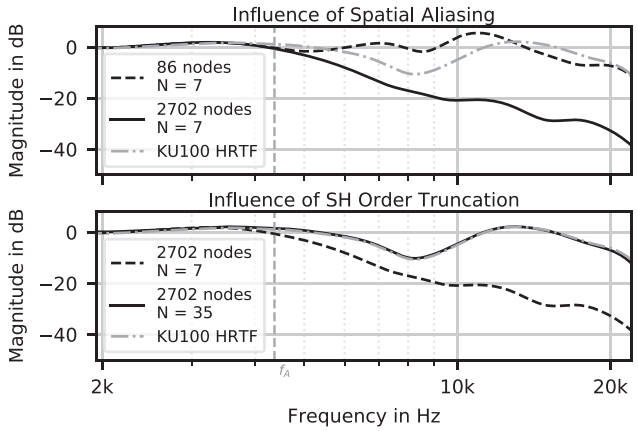


Fig. 2. Binaural signals obtained from the KU100 head-related impulse response (HRIR) set for a simulated plane wave impinging on simulated arrays of varying numbers of sampling nodes from the direction ($\phi = 0°,\ \theta = 90°$) with a maximum permitted radial filter gain of 40 dB. All curves are 1/3-octave-smoothed.

truncation of the SH coefficients leads to side lobes of the plane wave components from Eq. (2), which can also impair ILDs [12]. A side effect is the circumstance that the order truncation attenuates the signal at high time-frequencies to a considerable extent. This can be seem in Fig. 2 (bottom), where we used the HRIRs of the Neumann KU100 dummy head to calculate binaural signals according to Eq. (4) resulting from a simulated broadband plane wave impinging on virtual SMAs from ($\phi = 0°,\ \theta = 90°$). Both ear signals depicted in Fig. 2 (bottom) are based on a 2,702-node grid so that they exhibit a negligible amount of spatial aliasing. The attenuation of the magnitude is apparent at frequencies above 4 kHz.

Spatial aliasing constitutes spatial ambiguities, as information from higher modal orders appears in lower-order modes. This may likewise impair interaural cues. As a side effect, it results in an increase of the level at higher time-frequencies and therefore produces a high-shelf effect on the time-frequency response, as illustrated in Fig. 2 (top). The black curve depicts the left-ear binaural room transfer function (BRTF) based on a 2,702-node Lebedev grid SMA and thus contains no considerable spatial aliasing. The dashed curve is based on an 86-node grid and is affected by spatial aliasing that manifests in this representation as

an increase of the magnitude at frequencies above 4 kHz. Both signals were computed for the same SH order $N = 7$ to ensure identical truncation effects.

The left-ear measured KU100 HRIR is depicted in Fig. 2 for reference (grey dash-dotted curve). It can be seen that the SMA rendering up to $N = 35$ based on the 2,702 node grid (bottom) exactly matches the measured HRIR. The top figure shows that the high pass of spatial aliasing and the low-pass of order truncation cancel out each other. However, significant deviations from the reference persist.

1.3 Mitigation Approaches

A number of approaches to mitigate the impairment due to spatial undersampling in binaural rendering of SMA data have been presented in the past years. This section outlines the basic concepts of a selection of approaches. The same approaches are covered in the listening experiment that we conducted.

1.3.1 Bandwidth Extension Algorithm for Microphone Arrays

The Bandwidth Extension Algorithm for Microphone Arrays (BEMA) [13, 3] synthesizes the sound field SH coefficients of the higher time-frequency bands. It thus addresses the spatial ambiguities as well as the influence on the time-frequency transfer function. For this, spatial and spectral properties of the reliably obtainable frequency bands are acquired. The spatial energy distribution is extracted from the SH coefficients of frequency bands below the spatial aliasing frequency f_A as given by Eq. (5). The total energy of the higher frequencies is derived from an additional omnidirectional center microphone, ideally located in the center of the array. This approach is based on the observation that most relevant sound fields exhibit a smooth energy distribution in adjacent frequency bands.

The synthesis of the BEMA SH coefficients can be mathematically expressed as

$$S_{nm,\,\mathrm{BEMA}} = \underbrace{\frac{1}{d_n(\frac{\omega}{c}r)}\, I_{nm}}_{\text{spatial information}} \cdot \underbrace{C_0(\omega)}_{\text{spectral information}} \quad , \tag{7}$$

where C_0 is the energy normalized frequency domain signal of the center microphone and I_n^m the normalized so-called spatiotemporal image

$$I_{nm} = \frac{1}{W}\sum_{\mu=1}^{W} d_n\left(\frac{\omega_a - \mu}{c}r\right)\, S_{nm}(\omega_a - \mu) \tag{8}$$

with the averaging width W and the cut-off frequency ω_a, above which the BEMA synthesis is effective. The choice of W and ω_a defines the frequency bands denoted as source bands that are included in the calculation of I_{nm}.

1.3.2 Magnitude Least-Squares

Magnitude Least-Squares (MagLS) [14] is a method for reducing the impact of SH order truncation. This method premodifies the HRTF set in such a way that the energy in higher SH modes is reduced without notably decreasing the perceptual quality. If such higher modes are then removed due to truncation, the error becomes less significant. This modification is an advancement of the time alignment approach [15]. According to the duplex theory [16], interaural time differences become perceptually less important at high frequencies than ILDs. However, most of the energy in higher modes is caused by rapid phase changes towards higher frequencies. Thus, removing the linear phase at higher frequencies will decrease the energy in higher modes without significantly modifying the ILDs. The MagLS algorithm not only removes the linear phase slope but also minimizes the distance to the magnitudes of a reference HRTF set at higher frequencies. This is achieved by solving the least-squares problem in an iterative procedure according to

$$\min_{H_{nm}(\omega)} \sum_{q=1}^{Q} [\,|Y_n^m(\Omega_q)\,\mathbf{H}_{nm}(\omega)| - H(\Omega_q,\omega)]^2\,. \tag{9}$$

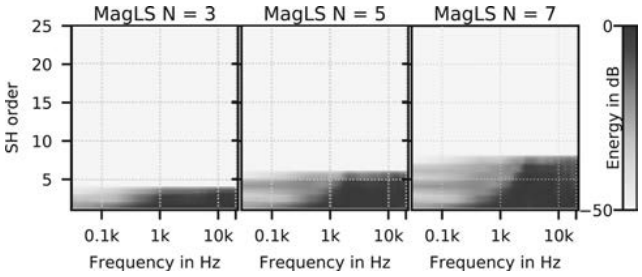


Fig. 3. Normalized logarithmic energy distribution of the HRTFs of the employed Neumann KU100 dummy head over frequency and SH order after MagLS preprocessing for the target orders $N = (3, 5, 7)$. The color encodes the energy ranging from −50 dB normalized to the maximum values for each frequency bin. It can be seen that MagLS modifies the information at low orders to account for the information that was removed from the higher orders.

The energy reduction in higher modes is depicted in Fig. 3. In contrast to the untreated HRTF set in Fig. 1, the energy in higher modes completely vanishes.

1.3.3 Spectral Equalization

To compensate for the modification of the time-frequency transfer function of the binaural signals, global equalization filters have been proposed. These filters are directly applied to the binaural signals and thus equalize every direction equally.

The so-called Spherical Head Filters (SHFs) [17] have been developed to compensate for the low-pass effect of SH order truncation. The authors determine the systematic magnitude deviation of order-truncated HRTFs based on a spherical head model and propose a global compensation filter without taking the effect of spatial aliasing into account. Applying these filters, which are depicted in Fig. 4, to all directions equally results in improved frequency responses for frontal directions but can make the deviations for lateral and especially contralateral sound incidents even larger.

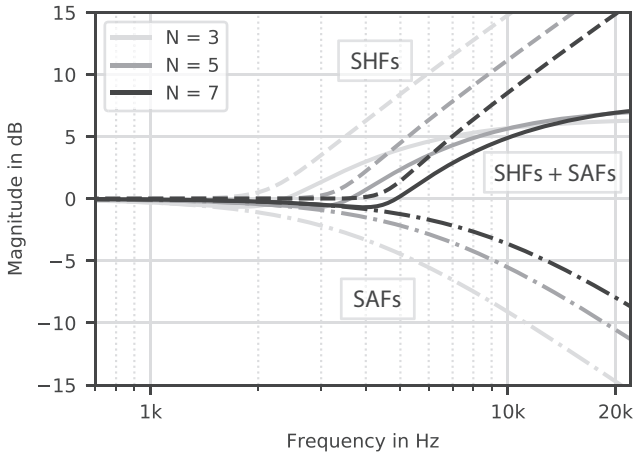


Fig. 4. Spherical Head Filters (dashed line), Spatial Aliasing Filter (dot-dashed line), and the combination of both (solid line) for orders $N = (3, 5, 7)$. Note that the Spherical Head Filters are designed with respect to the current SH rendering order N, the Spatial Aliasing Filters with respect to the maximum order N_{grid} that the sampling scheme permits. We assume $N = N_{\text{grid}}$ here.

An equalization approach to compensate for the high-shelf boost effect of spatial aliasing was proposed in [3]. The authors computed the deviation of dummy head measured room transfer functions to corresponding array renderings. For the array renderings, HRTFs with limited modal resolution were used to design the filters under negligible truncation errors [3, pp. 83], [18]. It was found that for diffuse sound fields, the average logarithmic deviations between dummy head transfer functions and array renderings follows a +6 dB/octave slope above f_{A}. Thus, aliasing compensation filters can be deployed generically using first order low-pass filters with the cut-off at f_{A}.

Informal listening showed that the low-pass effect of the truncation error is more noticeable than the high-shelf boost of spatial aliasing and solely applying the low-pass filter to compensate for aliasing yields no considerable perceptual benefit. We therefore combined the SHFs and the +6 dB/octave low-pass spatial aliasing filters (SAF), which results in a global undersampling equalization filter (SHF+SAF). Thus, we exclusively consider the SHFs and SHF+SAFs in the remainder.

1.3.4 Tapering

A method denoted as Spherical Harmonics Tapering to suppress the side lobes induced by order truncation was presented in [12]. Truncating the series of SH coefficients at a given order corresponds to applying a rectangular window over the order n, which results in considerable side lobes. The authors discussed different window functions and proposed a cosine-shaped fade-out towards higher orders as the most effective one. As any order-truncated signal, the resulting binaural signals need to be equalized by the previously discussed SHFs, whereby the Tapering requires slightly modified cut-off frequencies below f_{A}. In the remainder, we will solely discuss the combination of Tapering and SHF and denote this Tapering+SHF.

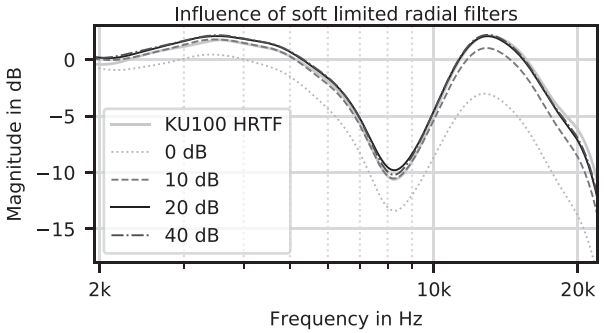


Fig. 5. Left ear magnitude responses of the frontal KU100 HRTF and ARIR binaural renderings up to 35th-order with different radial filter soft limits. The ARIR renderings are based on a simulated broadband plane wave impinging on an SMA with a 2,702-point Lebedev grid from ($\phi = 0°$, $\theta = 90°$). Our experiment employed the 0-dB limit.

2 EMPLOYED DATA

Many investigations are based on array room impulse responses (ARIRs) [19–21, 15, 4] as these allow for more flexibility regarding the design of the microphone array as well as more controlled conditions. Real-time implementations of the binaural rendering pipeline were presented e.g. in [1, 22, 23]. A noteworthy difference between ARIR-based rendering and the rendering of streamed (live-captured or recorded) signals is the fact that the signals from ARIR-based rendering are free from additive noise from the microphones and other stages in the signal chain, which can be strongly amplified by the radial filters d_n in (4). We employ a soft-limiting approach [3, pp. 90–118] that restricts the radial filter magnitudes to 0 dB. This was also done in the experiments in [3, 4] and may be considered to be on the conservative side so that the signal-to-noise ratio in the binaural signals is high even in the case that microphone self-noise and the like are apparent [24]. Fig. 5 illustrates the influence of the soft limiting. It shows the left ear BRTFs resulting from a broadband plane wave impact from ($\phi = 0°$, $\theta = 90°$) on a simulated 2,702-node Lebedev SMA. The BRTFs were calculated up to 35th-order using the different radial filter gain limits of 0, 10, 20, and 40 dB.

We used the `sound_field_analysis-py` Python toolbox [25] and the impulse response data set from [26] to prepare the stimuli. `sound_field_analysis-py` computes the radial filters $d_n(\omega)$ via sampling of the complex analytic frequency-domain representations resulting in impulse responses of length 2048 without time aliasing.

The impulse response data set contains BRIRs measured with a Neumann KU100 dummy head and ARIRs captured on various Lebedev grids under identical conditions. The ARIR measurements were performed with the VariSphear device [27], which is a fully automated robotic measurement system that sequentially captures directional impulse responses on a spherical grid for emulating a sphere microphone array. To obtain impulse responses of a rigid sphere array, an Earthworks M30 microphone was flush-mounted into a wooden spherical scattering body (see [26, Fig. 12]). All measurements were performed in four different rooms

at the WDR broadcast studios in Cologne, Germany. We employed the data sets of the rooms Small Broadcast Studio (SBS) and Large Broadcast Studio (LBS) with approximate reverberation times of 1 s and 1.8 s, respectively.

Binaural rendering of the ARIRs was performed according to Eq. (4) for a pure horizontal grid of head orientations with 1° resolution using the Neumann KU100 HRIR set, which were available on a 2,702-sampling-point Lebedev grid [28]. We denote these data "ARIR renderings" in the remainder. Likewise, the BRIRs of the same dummy head were available for the same head orientations so that a direct comparison of both auralizations was possible.

All mitigation algorithms were implemented with `sound_field_analysis-py`. Solely the MagLS HRIRs were preprocessed with MATLAB code provided by the authors of [14]. Every ARIR parameter (room, order, and sampling grid) set was processed with each of the algorithms MagLS, Tapering+SHF, SHF, and SHF+SAF. An untreated (Raw) ARIR rendering was also produced.

Previous experiments showed that SH representations of an order of less than 8 exhibit audible undersampling artifacts, i.e., a clear perceptual difference to the reference dummy head data was apparent [4]. As the present work investigates undersampled sound fields, we chose to focus on SH orders below 8 for the instrumental and perceptual evaluations as we cannot expect a considerable effect of the mitigation approaches for orders higher than that.

3 INSTRUMENTAL EVALUATION

In this section, the mitigation algorithms are evaluated and compared with a focus on their influence on the time-frequency spectrum. Fig. 6 depicts the logarithmic differences of left-ear BRTFs measured with a dummy head to BRTFs based on ARIR renderings of room SBS using the anechoic HRTF of that same dummy head. The left-hand plots are based on 50 sampling point grids rendered with order 3, the right-hand plots are based on 86 sampling point grids rendered with order 7. The vertical dashed lines indicate the spatial aliasing frequency f_A. The horizontal lines indicate the head orientation for which the rendered sound source is located contralateral to the depicted ear.

It can be seen that significant differences between dummy head BRTF and ARIR signals arise above f_A and especially for the contralateral direction for the Raw rendering, which does not employ any mitigation method.

The BEMA processed ARIR renderings exhibit considerably larger deviations. Even the authors of BEMA reported that the method introduces audible artifacts when applied to nonanechoic sound fields. As shown in [13], BEMA only works well for a single plane wave impact, whereas a low number of three phase-shifted plane waves impinging from different directions already leads to considerable comb filtering artifacts. Also, the averaging of the SH coefficients in the source band leads to a low-pass effect on the binaural signals.

Comparing SHF+SAF and SHF to the Raw condition shows that both equalizations reduce the spectral differences significantly. SHF+SAF exhibits slightly lower deviations than SHF, whereby both approaches still exhibit considerable deviations around the contralateral direction.

The ARIR renderings with SH Tapering exhibit similar spectral differences like the equalization approaches SHF+SAF and SHF. Recall that Tapering incorporates a modified SHF filter. Interestingly, although the modified SHFs were employed for the Tapering algorithm, the spectral differences are more similar to SHF+SAF.

Similarly to SHF, SHF+SAF, and Tapering+SHF, the MagLS processed ARIR renderings show significantly lower spectral differences than the Raw rendering. In the case of a 3rd-order rendering, MagLS clearly yields the result closest to the reference BRIR. For the more sophisticated SMA ($N = 7$), on the other hand, MagLS does not outperform the other approaches.

In summary, the instrumental evaluation shows that SHF, SHF+SAF, Tapering+SHF, and MagLS all reduce deviations of the time-frequency spectrum to a similar extent, whereas BEMA increases them. All methods cause deviations particularly for sources that are contralateral.

4 PERCEPTUAL EVALUATION

We conducted a listening experiment in order to examine to what extent the above introduced mitigation approaches provide perceptual improvements for the binaural rendering of undersampled SMA data. The subjects' task was to compare head-tracked auralizations of SMA data that were preprocessed with one of the mitigation methods to head-tracked auralizations of corresponding dummy head measurements.

4.1 Methods

4.1.1 Stimuli

The stimuli were generated for the SH orders 3, 5, and 7 as described in Sec. 2, for a pure horizontal grid with 1° resolution allowing for direct comparison of dummy head and ARIR auralizations. Informal pilot tests revealed that there are rather small audible differences of the mitigation methods for acoustically dry environments, we chose to use the data of the rooms SBS and LBS with exhibt reverberation times of 1 s or more (cf. Sec. 2). We used the ARIRs measured on the 50-sampling-point Lebedev grid for the ARIR renderings of SH order 3 and 5 and the 86-sampling-point Lebedev grid for order 7.

Previous studies showed a significant dependency of the perceived difference on the position of the auralized sound source [4, 15]. We therefore generated all ARIR renderings for two nominal head orientations: such that the virtual sound source appeared straight ahead ($\phi = 0°,\ \theta = 90°$), as well as such that it appeared lateral ($\phi = 90°,\ \theta = 90°$). Anechoic drum recordings were used as the test signal in particular because drums have a wide spectrum and strong transients, which makes them a critical test signal. Previous studies showed that certain aspects are only revealed with critical signals [3, 4]. To support transparency, static

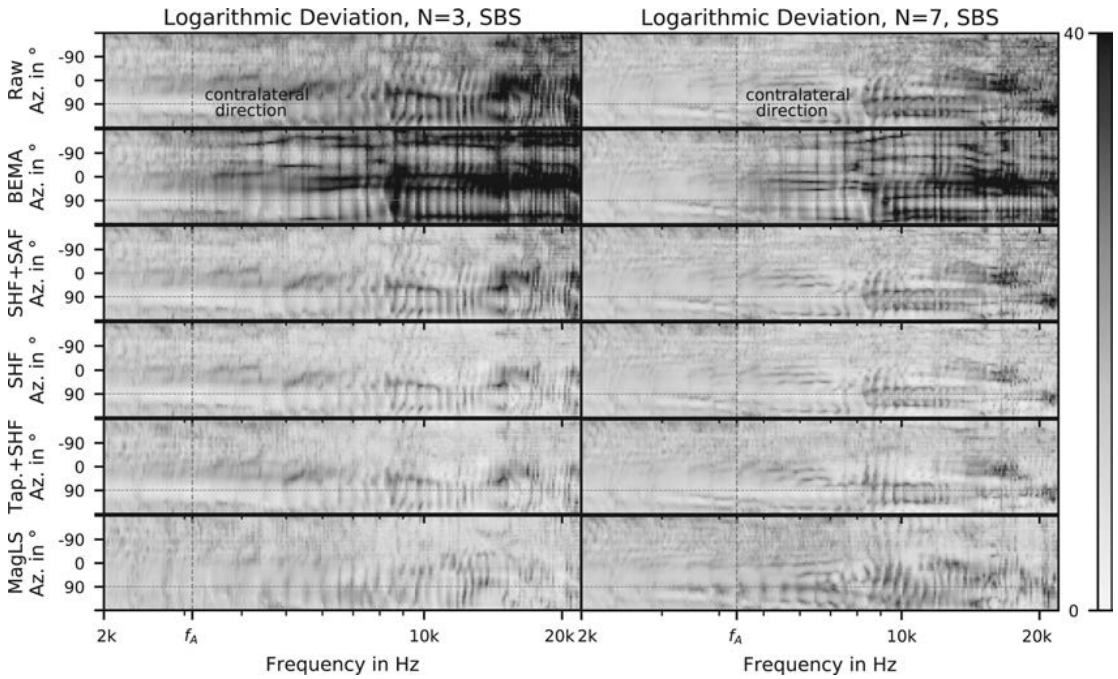


Fig. 6. Logarithmic deviations for the left ear of SBS ARIR renderings from the corresponding dummy head BRTFs with respect to azimuth angle of head orientation (vertical axes) and frequency (horizontal axes). The ARIR renderings were processed with each of the discussed algorithms. The shade of grey encodes the magnitude of the deviation ranging from 0–40 dB.

stimuli for both tested sound source positions are publicly available.[1]

4.1.2 Setup

The experiment was conducted in a quiet acoustically damped audio laboratory at Chalmers University of Technology. The SoundScape Renderer (SSR) [29, 30] in binaural room synthesis (BRS) mode was used for dynamic auralization. It convolves arbitrary input test signals with a pair of BRIRs corresponding to the instantaneous head orientation of the listener, which was tracked along the azimuth with a Polhemus Patriot tracker. A change of head orientation as well as switching between stimuli results in a cross-fade with cosine ramps over the course of one processing block. All stimuli were time aligned so that no artifacts occurred during the fade.

The binaural renderings were presented to the participants using AKG K702 headphones with a Lake People G109 headphone amplifier at a playback level of about 66 dBA. The output signals of the SSR were routed to an Antelope Audio Orion 32 DA converter at 48 kHz sampling frequency and a buffer length of 512 samples. Equalization according to [26] was applied to compensate for the headphone transfer function. All involved software components were running on the same iMac Pro 1.1 computer.

4.1.3 Paradigm

We used a test design based on the Multiple Stimulus with Hidden Reference and Anchor (MUSHRA) method as proposed by the International Telecommunication Union (ITU) [31]. The subjects' task was to rate the overall perceived difference between the ARIRs renderings and the dummy head reference. We used a non-head-tracked diotic 0° dummy head reference BRIR of the room under test as anchor, which was low-pass filtered with a cutoff at 3 kHz.

Table 1. The stimuli employed in the listening experiment. All algorithms were presented in each of the trials. Each such set was rendered for 3 SH orders, 2 source positions, and 2 rooms. This results in 12 trials with 8 stimuli each.

Algorithm	SH order (grid)	Position	Room
BEMA	3 (50)	$\phi = 0°, \theta = 90°$	LBS
MagLS	5 (50)	$\phi = 90°, \theta = 90°$	SBS
SHF	7 (86)		
SHF + SAF			
Tapering + SHF			
Raw			
Reference			
Anchor			

Each trial required 8 ratings to be performed by the subject (BEMA, MagLS, SHF, Tapering, SHF+SAF, Raw, hidden reference, anchor) against the dummy head reference. The experiment consisted of 12 trials: 3 SH orders (3, 5, 7) × 2 nominal source positions (0°, 90°) × 2 rooms (LBS, SBS), as summarized in Tab. 1. The subjects were provided with a graphical user interface (GUI) with continuous sliders named as "No difference" (100), "Small difference" (75), "Moderate difference" (50), "Significant difference" (25), and "Huge difference" (0) as depicted in Fig. 7.

[1] https://doi.org/10.5281/zenodo.3759343

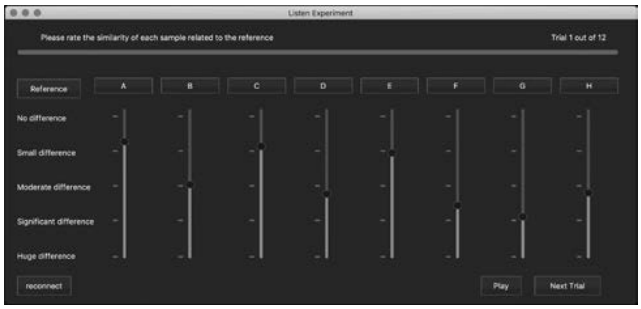


Fig. 7. PyQt GUI used in the listening experiment.

4.1.4 Procedure

Twenty participants, 4 of them female, between the ages of 22 and 50 took part in the experiment. Most of them were M.Sc. students or staff at the Division of Applied Acoustics of Chalmers University of Technology. Sixteen participants reported that they had previously participated in a listening experiment. The subjects were sitting in front of a computer screen with a keyboard and a mouse. It was possible to listen to each stimulus as often and long as desired. The participants were allowed and strongly encouraged to move their heads during the presentation of the stimuli. At the beginning of each experiment, the subjects had to rate four training stimuli that covered a representative set of perceptual differences of the presented stimuli in the subsequent test. These training stimuli consisted of a BEMA and MagLS rendering of SBS data at 3rd order for the lateral sound source position, as well as the corresponding anchor and reference. The experiment took on average about 40 minutes per participant.

4.2 Results

All anchor and reference ratings were post-screened before applying statistical analysis according to the recommendation of the ITU [31]. All anchor ratings were below 30 and most reference ratings above 80. Only two reference ratings (50, 49) and two anchor ratings (40, 38) were conspicuous, which constitutes a low portion of in total 96 ratings per participant. We performed statistical analyses including and excluding the respective subjects' data, which led to identical results. We report only the results over the complete data set here.

Our subjects compared the different algorithms for a given combination of SH order, room, and source position in each trial. We therefore highlight that a direct comparison of the ratings for different orders, rooms, and source positions has to be performed with reservation. The anchors and references were conceptually the same across all trials and the stimulus and condition order were randomized per participant. A certain amount of consistency in the subjects'

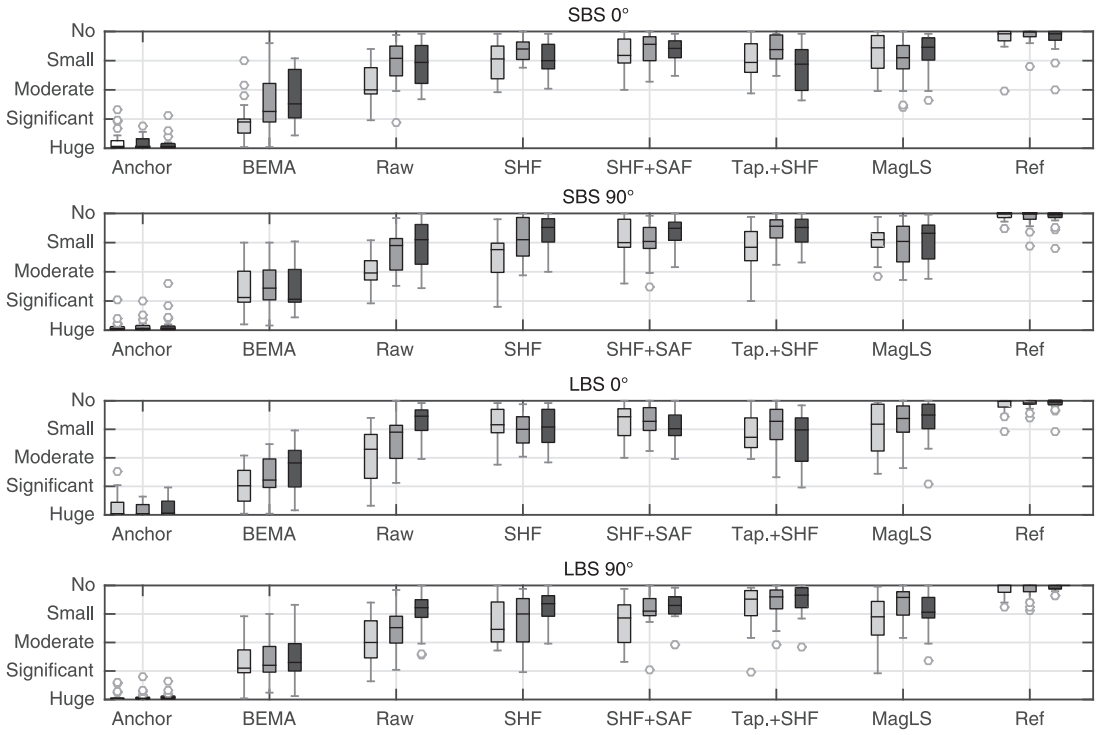


Fig. 8. Boxplots illustrating the ratings of the perceptual difference between the stimulus and the dummy head reference for each room and virtual source position separately. Each figure depicts the boxplots for each algorithm at the SH order 3 (light grey), 5 (dark grey), and 7 (black), respectively. Each box indicates the 25th and 75th percentiles, the median value (black line), the outliers (grey circles), and the minimum/maximum ratings not identified as outliers (black whiskers).

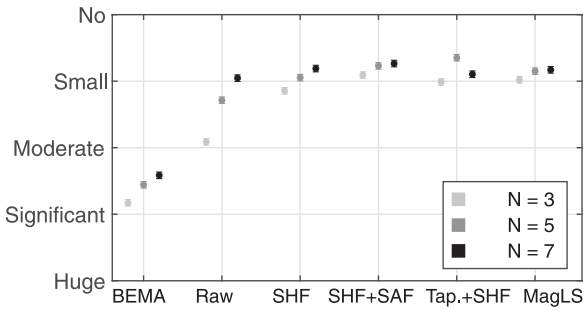


Fig. 9. Mean difference ratings pooled over source position and room with respect to algorithm (abscissa) and SH order (colors). The 95% within-subject confidence intervals were calculated according to [33, 34].

responses may therefore be assumed. In the following, we present a statistical analysis that includes the comparison between orders and positions as it is commonly performed with MUSHRA data. This facilitates discussing the results in relation to the literature as we will do in Sec. 4.3.

An overview of the results is presented as boxplots in Fig. 8, illustrating the ratings for the rooms SBS and LBS and source positions at 0° and 90° separately. The boxplots confirm that subjects rated the hidden anchor and the reference consistently. Furthermore, perceptual differences between Raw and dummy head renderings tended to become smaller with increasing SH order. All algorithms with the exception of BEMA led to a smaller perceptual difference to the reference than the Raw renderings.

For statistical analysis of the results, repeated measures ANOVAs were performed. A Lilliefors test for normality was applied to test the requirements for the ANOVA. It failed to reject the null hypothesis in 14 of 72 conditions at a significance level of $p = 0.05$. However, parametric tests such as the ANOVA are generally robust to violations of normality assumption [32]. For further analysis, Greenhouse–Geisser-corrected p values are considered, with the associated ϵ-values for correction of the degrees of freedom of the F-distribution being reported.

A four-way repeated measures ANOVA with the within-subject factors algorithm (BEMA, MagLS, Tapering+SHF, SHF, SHF+SAF, and Raw), order (3, 5, 7), room (SBS, LBS), and nominal source position (0°, 90°) was performed. In addition, a number of nested repeated measures ANOVAs were performed. For each of the six algorithms, one three-way ANOVA with the factors order (3, 5, 7), room (SBS, LBS), and source position (0°, 90°), as well as a four-way ANOVA with the subset of MagLS, Tapering+SHF, SHF, and SHF+SAF for the factor algorithm and the factors order (3, 5, 7), room (SBS, LBS), and source position (0°, 90°) were applied. The results of the ANOVA incorporating all algorithms are presented in Tab. 2.

The results of the experiment are depicted in aggregate form in Fig. 9. The mean values with respect to algorithm and SH order are depicted separately. Each value was calculated by averaging the ratings of all participants, source positions and rooms. Furthermore, 95% within-subject confidence intervals as proposed by [33, 34], based on the main effect of the algorithm, are shown. The plots confirm the observations taken from the boxplots and additionally show that the ratings do not scale linearly with the rendering order. It is noteworthy that on average, the Tapering renderings were rated with a larger perceptual difference when rendered at SH order 7 than with order 5.

Overall, the ratings of the algorithms SHF, SHF+SAF, Tapering+SHF, and MagLS are located in a similar range. We therefore preliminary conclude that all algorithms achieve a similar magnitude of improvement compared to Raw renderings.

The following analysis refers to main effects and first order interactions only. It was found for the four-way ANOVA involving all algorithms that the algorithm and order main effects as well as the first order interaction effects algorithm×order, algorithm×position, and order×position were significant. These effects will be examined successively in the following paragraphs.

The main effects of the algorithm ($F(5, 95) = 194.9$, $p < .001$, $\eta_p^2 = .911$, $\epsilon = .684$) as well as of the order ($F(2, 38) = 40.75$, $p < .001$, $\eta_p^2 = .682$, $\epsilon = .765$) support the trends identified in the boxplots in Fig. 8 and the mean plots in Fig. 9. All algorithms significantly affect the perceived similarity and for all algorithms other than Tapering+SHF, higher SH orders yield more perceived similarity. Furthermore, the interaction effect algorithm×order ($F(10, 190) = 8.06$, $p < .001$, $\eta_p^2 = .298$, $\epsilon = .612$) suggests that both factors do not just exclusively influence the perceived differences, but the algorithms may lead to different levels of improvements with respect to the SH order.

To validate the observation that the algorithms SHF, SHF+SAF, Tapering+SHF, and MagLS achieved similar improvements, a four-way repeated measures ANOVA was performed taking into account only the results for these algorithms. The factor algorithm was not significant ($p = .107$), showing that all algorithms except BEMA achieved similar perceptual improvements. The ANOVAs conducted for each algorithm separately suggested no significant effect for the SH order for the algorithm MagLS only ($p = .202$). This indicates that MagLS performs comparably similar at all orders.

We found no main effect of the factor source position ($p = .49$), but an interaction effect of algorithm×position ($F(5, 95) = 5.563$, $p < .001$, $\eta_p^2 = .227$, $\epsilon = .001$). This suggests that the algorithms perform differently dependent on the presented source position. Moreover, the ANOVA revealed a significant interaction of order×position ($F(2, 38) = 194.9$, $p < .026$, $\eta_p^2 = .187$, $\epsilon = .858$). Thus, the position dependency varies with respect to the order. The results of the ANOVA can be seen in Fig. 10 (left, right), presenting the mean values calculated similarly to Fig. 9 but separated into frontal and lateral nominal source position. The plots indicate that the 7th-order renderings processed with the SHF, SHF+SAF, and Tapering+SHF algorithms were rated with larger perceptual difference for frontal than for lateral sound source positions.

Interestingly, an ANOVA over exclusively the data of any one given algorithm suggests that Tapering+SHF is the only single algorithm for which a significant main effect of the source position ($F(1, 19) = 15.61$, $p < .001$, $\eta_p^2 = .451$,

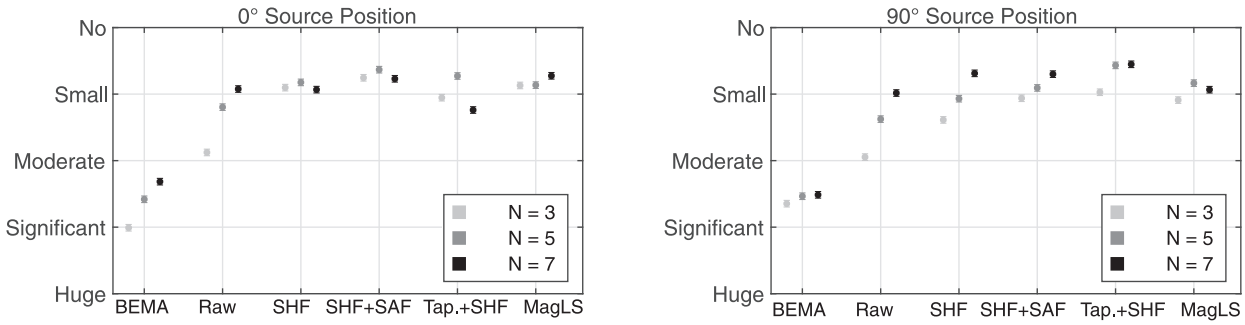


Fig. 10. Mean difference ratings for 0° (left) and 90° (right) sound source position pooled over both rooms with respect to algorithm (abscissa) and SH order (colors). The 95% within-subject confidence intervals were calculated according to [33, 34].

$\epsilon = 1$) may be apparent. To further dissect this observation, we performed multiple t tests (with Hochberg correction to correct for multiple hypothesis testing), comparing 5th and 7th-order renderings processed with SHF, SHF+SAF, and Tapering+SHF for frontal and lateral source positions. The tests suggest a significant difference between the ratings for frontal and lateral source position only for 7th-order renderings with Tapering+SHF ($t(39) = 4.879$, $p < .001$, $d_z = .772$). This indicates a rather weak influence of source position and order in the present data set. Concerning the influence of the room, we found neither a main effect nor any interaction effect.

4.3 Discussion

The results of the perceptual evaluation show that all presented algorithms other than the BEMA approach yield perceivable improvements of binaural array renderings. No algorithm was rated significantly better than the others. All analysis of the dependency of the ratings on the rendering order, room, and source position has to be performed with reservation as this requires comparing ratings across different trials. As we argued in Sec. 4.2, a considerable amount of consistency of the ratings may be assumed across trials. We discuss our data in relation to findings from similar studies and analyses in the literature in the following.

Our listening experiment confirms that higher-order renderings were mostly rated closer to the dummy head than lower order renderings. However, all orders we tested were rated significantly different compared to the reference. This matches the findings from [3, 4] where it was found that renderings of an order below 8 exhibit audible differences to dummy head auralizations and that these differences are induced by spatial undersampling. The soft limiting that we applied to the radial filters may have led to audible differences of ARIR and dummy head auralization independent of undersampling. This may have caused a saturation of the perceptual improvement towards higher orders. Similar results were obtained in [3, 4] where similar soft limiting was applied. We assume that less-conservative radial filter limits lead to more similarity to the dummy head in particular at higher frequencies, as indicated by Fig. 5. The cost is a lower signal-to-noise ratio in the binaural signals if additive sensor noise is apparent [24]. Similarly to [3, 4], we observed no room dependency in the ratings.

Table 2. Results of the four-way repeated measures ANOVA with the within-subject factors algorithm (BEMA, MagLS, Tapering+SHF, SHF, SHF+SAF, Raw), order (3, 5, 7), source position (0°, 90°), and room (SBS, LBS).

Effect	df	F	ϵ_{GG}	η_p^2	p_{GG}
Algorithm	5	194.898	.684	.911	<.001*
Order	2	40.750	.765	.682	<.001*
Position	1	.495	1.000	.025	.490
Room	1	1.617	1.000	.078	.219
Algorithm×Order	10	8.055	.612	.298	<.001*
Algorithm×Position	5	5.565	.653	.227	.001*
Order×Position	2	4.372	.858	.187	.026*
Algorithm×Room	5	1.001	.742	.050	.409
Order×Room	2	.731	.709	.037	.446
Position×Room	1	.181	1.000	.009	.676
Algorithm×Order×Position	10	4.479	.644	.191	<.001*
Algorithm×Order×Room	10	3.218	.549	.145	.008*
Algorithm×Position×Room	5	1.445	.713	.071	.233
Order×Position×Room	2	.478	.913	.025	.607
Algorithm×Order×Position×Room	10	.909	.622	.046	.494

ϵ_{GG}: Greenhouse–Geisser epsilons
η_p^2: Partial eta squared
p_{GG}: Greenhouse–Geisser corrected p values
Statistical significance at 5% level are indicated by asterisks

Previous studies observed a dependency of the ratings on the sound source position. In [4], the participants compared dummy head auralizations and raw ARIR renderings in terms of spaciousness and timbre separately. The timbre of ARIR renderings of SH orders of 8 and higher was perceived noticeably closer to the dummy head auralization for lateral sound sources than for frontal sources, which at first glance seems surprising considering the deviations of truncated SH representations at the contralateral ear for lateral source positions (cf. Fig. 6). The authors concluded that spectral differences of frontal sound sources can be perceived more reliably than the spectral differences of lateral sources and attributed it to the higher spatial resolution of the human auditory system in the front [35].

In contrast, another study presented in [15] showed that ARIR renderings treated with the time alignment approach (a predecessor of MagLS) were rated lower for lateral than for frontal sources. However, even though the plots of our results in Fig. 10 indicate some amount of source position dependency, the statistical analysis revealed a significant effect of source position only for the algorithm Tapering+SHF. There is therefore no statistical evidence in our data that supports the observations of a general influence of source position on difference ratings that was made in [4, 15].

The statistically significant effect of order and source position for the algorithm Tapering+SHF, e.g., that 5th-order renderings were sometimes rated higher than 7th-order renderings, might have been caused by unfavorable effects of the tapering with higher-order data. The higher the rendering order, the more SH modes are attenuated by the window, as the tapering starts always at $n = 1$ independent of the maximum order. Tapering in its present form might therefore be most beneficial for rendering orders below 7. A modified approach that tapers only the last few orders below the maximum order is conceivable.

As discussed above, these observation can ultimately only be proven based on data from a direct comparison of stimuli of different orders.

5 CONCLUSIONS

A listening experiment comparing different algorithms to mitigate the perceptual impairment of binaural rendering of SMA data due to spatial undersampling was presented. The subjects' task was to compare array renderings enhanced with state-of-the-art algorithms to corresponding auralizations of dummy head impulse response data in terms of overall perceived difference.

We found that the Magnitude Least-Squares HRIR preprocessing approach, the Spherical Head Filters, and the Spherical Harmonics Tapering (including the SHF), as well as a global undersampling equalization filter, all yield a significant improvement of the SMA renderings. We only evaluated the overall perceived difference to the dummy head auralization and can therefore not break down the differences into individual attributes. A follow-up study, evaluating these attributes such as spaciousness or timbre separately may expose individual advantages and disadvantages of the investigated approaches in more detail. It may be assumed that appropriate equalization of the spectrum yields improvements in particular for the timbre, whereas MagLS and Tapering may be more beneficial for improving the localization and thus, the spaciousness of the binaural synthesis.

The Bandwidth Extension Algorithm for Microphone Arrays is the only algorithm aiming at recovering the loss of spatial information due to spatial aliasing that seemed to produce more harm than benefit. This is mainly caused by the low-pass effect of the involved SH coefficient averaging. Magnitude Least-Squares and Tapering have shown to be appropriate algorithms to mitigate the truncation artifacts, but also, a simple equalization of the binaural time-frequency response, i.e., the Spherical Head Filters and the global undersampling equalization filter, yielded perceptually equivalent results. Simple (global) equalization has the disadvantage of shifting coloration impairments, by design, mostly to the contralateral side.

Although most tested algorithms are successful in improving the array auralizations, there are still audible differences to the corresponding dummy head reference. These differences may be related to spatial ambiguities of spatial aliasing. Instrumental analysis as well as informal listening revealed that the modification of the time-frequency response is more affected by SH order truncation than by spatial aliasing. It remains to be clarified whether the overall perceptual influence of truncation is more significant, and whether spatial aliasing artifacts can even be neglected for sufficient auralizations.

Some amount of the saturation of the observed perceived differences may also be attributed to the choice of radial filter limit, which caused a slight attenuation of the signal at higher frequencies.

6 ACKNOWLEDGMENT

We thank all participants of the listening experiment for their voluntary support and Christian Schörkhuber, Markus Zaunschirm, and Franz Zotter of Institute of Electronic Music and Acoustics at the University of Music and Performing Arts in Graz for providing us with their code of MagLS. We also thank all anonymous reviewers for very valuable and stimulating feedback.

7 REFERENCES

[1] F. Zotter and M. Frank, *Ambisonics A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality* (Springer-Verlag, Berlin, Germany, 2019), doi:10.1007/978-3-030-17207-7, https://plugins.iem.at/.

[2] J. Ahrens, *Analytic Methods of Sound Field Synthesis* (Springer, Berlin, Germany, 2012), doi:10.1007/978-3-642-25743-8.

[3] B. Bernschütz, *Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording*, Ph.D. thesis, Technische Universität Berlin (2016), doi:10.14279/depositonce-5082.

[4] J. Ahrens and C. Andersson, "Perceptual Evaluation of Headphone Auralization of Rooms Captured With Spherical Microphone Arrays With Respect to Spaciousness and Timbre," *Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2783–2794 (2019 Apr.), doi:10.1121/1.5096164.

[5] B. Rafaely, *Springer Topics in Signal Processing Springer Topics in Signal Processing* (Springer, Berlin, Germany, 2015), doi:10.1007/978-3-642-11130-3.

[6] E. G. Williams, *Fourier Acoustics* (Academic Press, London, United Kingdom, 1999), doi:10.1016/B978-0-12-753960-7.X5000-1.

[7] C. Andersson, *Headphone Auralization of Acoustic Spaces Recorded with Spherical Microphone Arrays*, Master's thesis, Chalmers University of Technology (2017).

[8] S. Lösler and F. Zotter, "Comprehensive Radial Filter Design for Practical Higher-Order Ambisonic Recording," presented at the *41st DAGA*, 1, pp. 452–455 (2015), doi:10.3758/BF03210951.

[9] C. E. Shannon, "Communication in the Presence of Noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21 (1949), doi:10.1109/JRPROC.1949.232969.

[10] B. Rafaely, "Analysis and Design of Spherical Microphone Arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 135–143 (2005 Jan.), doi:10.1109/TSA.2004.839244.

[11] Z. Ben-Hur, D. L. Alon, B. Rafaely, and R. Mehra, "Loudness Stability of Binaural Sound With Spherical Harmonic Representation of Sparse Head-Related Transfer Functions," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1 (2019 Dec.), doi:10.1186/s13636-019-0148-x.

[12] C. Hold, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev, "Improving Binaural Ambisonics Decoding by Spherical Harmonics Domain Tapering and Coloration Compensation," presented at the *International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 261–265 (2019), doi:10.1109/ICASSP.2014.6854442.

[13] B. Bernschütz, "Bandwidth Extension for Microphone Arrays," presented at the *133rd Convention of the Audio Engineering Society* (2012 Oct.), convention paper 8751.

[14] C. Schörkhuber, M. Zaunschirm, and R. Holdrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," presented at the *44th DAGA*, 4, pp. 339–342 (2018).

[15] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural Rendering of Ambisonic Signals by Head-Related Impulse Response Time Alignment and a Diffuseness Constraint," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3616–3627 (2018 Jun.), doi:10.1121/1.5040489.

[16] L. Rayleigh, "XII. On Our Perception of Sound Direction," *Philosophical Magazine Series 6*, vol. 13, no. 74, pp. 214–232 (1907), doi:10.1080/14786440709463595.

[17] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral Equalization in Binaural Signals Represented by Order-Truncated Spherical Harmonics," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4087–4096 (2017 Jun.), doi:10.1121/1.4983652.

[18] B. Bernschütz, A. Vázquez Giner, C. Pörschmann, and J. M. Arend, "Binaural Reproduction of Plane Waves With Reduced Modal Order," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 972–983 (2014 Oct.), doi:10.3813/AAA.918777.

[19] B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl, "SOFiA Sound Field Analysis Toolbox," presented at the *International Conference on Spatial Audio (ICSA)*, pp. 8–16 (2011 Jan.).

[20] J. Sheaffer, S. Villeval, and B. Rafaely, "Rendering Binaural Room Impulse Responses From Spherical Microphone Array Recordings Using Timbre Correction," presented at the *Joint Symposium on Auralization and Ambisonics (EAA)*, pp. 3–5 (2014 Apr.).

[21] A. Neidhardt, *Untersuchungen zur räumlichen Genauigkeit bei der binauralen Auralisation von Kugelarraydaten*, Master's thesis, Technische Universität Graz (2015).

[22] H. Helmholz, C. Andersson, and J. Ahrens, "Real-Time Implementation of Binaural Rendering of High-Order Spherical Microphone Array Signals," presented at the *45th DAGA*, pp. 2–5 (2019), https://github.com/AppliedAcousticsChalmers/ReTiSAR.

[23] L. Mccormack and A. Politis, "SPARTA & COMPASS: Real-Time Implementations of Linear and Parametric Spatial Audio Reproduction and Processing Methods," presented at the *AES Conference on Immersive and Interactive Audio* (2019 Mar.), conference paper 111, doi:10.1121/1.3278605, https://github.com/leomccormack/SPARTA.

[24] H. Helmholz, D. L. Alon, S. V. A. Garí, and J. Ahrens, "Instrumental Evaluation of Sensor Self-Noise in Binaural Rendering of Spherical Microphone Array Signals," presented at the *Forum Acusticum*, pp. 1–8 (2020).

[25] C. Hohnerlein and J. Ahrens, "Spherical Microphone Array Processing in Python With the Sound Field Analysis-Py Toolbox," *Fortschritte der Akustik – DAGA 2017*, pp. 1033–1036 (2017).

[26] P. Stade, B. Bernschütz, and M. Rühl, "A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios," presented at the *27th Tonmeistertagung - VDT International Convention*, pp. 551–567 (2012 Nov.).

[27] B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl, "Entwurf und Aufbau eines variablen sphärischen Mikrofonarrays für Forschungsanwendungen in Raumakustik und Virtual Audio," presented at the *36th DAGA*, pp. 717–718 (2010).

[28] B. Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," presented at the *39th DAGA*, pp. 592–595 (2013).

[29] M. Geier, J. Ahrens, and S. Spors, "The Soundscape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods," presented at the

124th Convention of the Audio Engineering Society (2008 May), convention paper 7330.

[30] M. Geier, J. Ahrens, and S. Spors, "The SoundScape Renderer," http://spatialaudio.net/ssr/ (2019), version 0.4.2, retrieved 2019-07-01.

[31] ITU-R BS.1534-3, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems" (2015).

[32] J. Bortz and C. Schuster, *Statistik für Human- und Sozialwissenschaftler*, 7th ed. (Springer-Verlag, Gießen, Germany, 2010), doi:10.1121/1.3278605.

[33] G. R. Loftus, "Using Confidence Intervals in Within-Subject Designs," *Psychonomic Bulletin & Review*, vol. 1, no. 4, pp. 1–15 (1994), doi:10.3758/BF03210951.

[34] J. Jarmasz and J. G. Hollands, "Confidence Intervals in Repeated-Measures Designs: The Number of Observations Principle," *Canadian Journal of Experimental Psychology*, vol. 63, no. 2, pp. 124–138 (2009 Jun.), doi:10.1037/a0014164.

[35] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, Massachusetts, 1997).

THE AUTHORS


Tim Lübeck


Hannes Helmholz


Johannes M. Arend


Christoph Pörschmann


Jens Ahrens

Tim Lübeck received his B.Sc. degree in Electrical Engineering in 2017 and his M.Sc. degree in Communication Engineering in 2019 from TH Köln, Cologne, Germany. He completed his master's thesis in cooperation with the Division of Applied Acoustics at Chalmers University. Since 2019, he has been a Research Fellow and working toward a Ph.D. at TH Köln and TU Berlin in the field of virtual acoustics, binaural technology, auditory perception, and audio signal processing.

•

Hannes Helmholz received his B.Sc. degree in Applied Computer Science from the University of Applied Sciences, Berlin, Germany, in 2013. and his M.Sc. degree in Audio Communication and Technology from Technische Universität Berlin, Berlin, Germany, in 2018. Since 2018, he has been working toward a Ph.D. in Applied Acoustics at Chalmers University of Technology, Gothenburg, Sweden, in the field of spatial audio, binaural technology, auditory perception, and audio signal processing.

•

Johannes M. Arend received his B.Eng. degree in media technology from HS Düsseldorf, Düsseldorf, Germany, in 2011 and his M.Sc. degree in media technology from TH Köln, Cologne, Germany, in 2014. Since 2015, he has been a Research Fellow and working toward a Ph.D. at TH Köln and TU Berlin in the field of spatial audio with a focus on binaural technology, auditory perception, and audio signal processing.

•

Christoph Pörschmann studied Electrical Engineering at the Ruhr-Universität Bochum (Germany) and at Uppsala Universitet (Sweden). In 2001, he obtained his Doctoral Degree (Dr.-Ing.) from the Electrical Engineering and Information Technology Faculty of the Ruhr-Universität Bochum as a result of his research at the Institute of Communication Acoustics. Since 2004, he has been Professor of Acoustics at TH Köln – University of Applied Sciences. His research interests are in the fields of virtual acoustics, spatial hearing, and the related perceptual processes.

•

Jens Ahrens has been an Associate Professor within the Division of Applied Acoustics at Chalmers University since 2016. He has also been a Visiting Professor at the Applied Psychoacoustics Lab at the University of Huddersfield, Huddersfield, UK, since 2018. Jens received his Diploma (equivalent to a M.Sc.) in Electrical Engineering/Sound Engineering jointly from Graz University of Technology and the University of Music and Dramatic Arts, Graz, Austria, in 2005. He completed his Doctoral Degree (Dr.-Ing.) at the Technische Universität Berlin, Berlin, Germany, in 2010. From 2011 to 2013, Jens was a Postdoctoral Researcher at Microsoft Research in Redmond, Washington, USA, and in the fall and winter terms of 2015/2016, he was a Visiting Scholar at the Center for Computer Research in Music and Acoustics (CCRMA) at Stanford University, Stanford, California, USA. He is an Associate Editor of the IEEE Signal Processing Letters and a Guest Editor of the EURASIP Journal on Audio, Speech, and Music Processing.

6 Voice Directivity

6.1 Investigating Phoneme-Dependencies of Spherical Voice Directivity Patterns

Pörschmann, C., & Arend, J. M. (2021). *J. Acoust. Soc. Am.*, *149*(6), 4553–4564. https://doi.org/10.1121/10.0005401

JASA ARTICLE

Investigating phoneme-dependencies of spherical voice directivity patterns[a)]

Christoph Pörschmann[b)] and Johannes M. Arend[c)]
Institute of Communications Engineering, TH Köln–University of Applied Sciences, Betzdorfer Str. 2, 50679 Cologne, Germany

ABSTRACT:
Dynamic directivity is a specific characteristic of the human voice, showing time-dependent variations while speaking or singing. To study and model the human voice's articulation-dependencies and provide datasets that can be applied in virtual acoustic environments, full-spherical voice directivity measurements were carried out for 13 persons while articulating eight phonemes. Since it is nearly impossible for subjects to repeat exactly the same articulation numerous times, the sound radiation was captured simultaneously using a surrounding spherical microphone array with 32 microphones and then subsequently spatially upsampled to a dense sampling grid. Based on these dense directivity patterns, the spherical voice directivity was studied for different phonemes, and phoneme-dependent variations were analyzed. The differences between the phonemes can, to some extent, be explained by articulation-dependent properties, e.g., the mouth opening size. The directivity index, averaged across all subjects, varied by a maximum of 3 dB between any of the vowels or fricatives, and statistical analysis showed that these phoneme-dependent differences are significant. © *2021 Acoustical Society of America*. https://doi.org/10.1121/10.0005401

(Received 4 February 2021; revised 28 May 2021; accepted 1 June 2021; published online 28 June 2021)
[Editor: Vasileios Chatziioannou] Pages: 4553–4564

I. INTRODUCTION

The directional properties of human voice radiation have been investigated for more than 200 years. Saunders (1790) determined the distance from which a listener can hear a human speaker for different radiation directions along the azimuth plane, resulting in an implicit description of human voice directivity. He applied these results to design suitable shapes of theaters and suggested optimal distances between the actor on stage and the audience. Similarly, Wyatt (1813) and Henry (1857) carried out scientific studies on human voice radiation to optimize theaters and lecture halls. These studies, which are summarized in Postma *et al.* (2018) in greater detail, show that the influence of human voice directivity and its effects on speech intelligibility have been understood for a long time. Trendelenburg (1929) performed the first direct measurements of human voice radiation, determining directivity patterns for several vowels and fricatives in the horizontal plane. Ten years later, Dunn and Farnsworth (1939) published comprehensive results analyzing the sound radiation for a spoken sentence in octave or third-octave bands from 63 Hz to 12 kHz. The authors measured the directivity along a sphere in steps of 45° in the horizontal and the vertical planes. The measurements were performed sequentially by comparing the microphone signals at the respective positions to a measurement with another microphone at a reference position. The data were measured eight times for different distances and then averaged, with the closest measurements made directly at the mouth opening and the farthest at a distance of 1 m. Due to technical restrictions, the authors had to perform the measurements for each frequency band separately, resulting in about 5000 measurements.

Several authors investigated the specific directional characteristics of the singing voice. Marshall and Meyer (1985) determined the directivity in both the horizontal and vertical planes. The results showed influences of voice dynamics on directivity, with differences of up to 3 dB between forte and piano and some slight differences between males and females. However, no statistical analysis was performed, as the study only relied on three subjects. For spoken utterances, these results were mainly confirmed by Chu and Warnock (2002), who also observed significant differences depending on the articulation level. Comparing the directivity while speaking or singing, Monson *et al.* (2012) did not find significant differences, only some slight dependencies on the articulation level. Both Chu and Warnock (2002) and Monson *et al.* (2012) did not find significant differences between males and females in their studies. Based on horizontal plane measurements with eight professional opera singers in three different rooms, Cabrera *et al.* (2011) determined substantial variations of voice directivity among singers and between singing in different rooms. Katz and D'Alessandro (2007) analyzed directivity patterns in the horizontal plane for sustained vowels articulated by a professional opera singer. This study showed no systematic differences for the different vowels. In contrast, Marshall and Meyer (1985) analyzed sound radiation for

[a)]This paper is part of a special issue on Modeling of Musical Instruments.
[b)]Electronic mail: christoph.poerschmann@th-koeln.de, ORCID: 0000-0003-0794-0444.
[c)]Also at: Audio Communication Group, Technical University of Berlin, Einsteinufer 17c, D-10587 Berlin, Germany. Electronic mail: johannes.ar-end@th-koeln.de, ORCID: 0000-0002-5403-4076.

different vowels and found differences between a sustained [e] and an [o] or a [u]. Confirming this study, Kocon and Monson (2018) examined articulation-dependent effects of voice radiation from different vowels and found the strongest directivity for an [a]. Monson *et al.* (2012) analyzed time-variant effects of the directivity pattern with measurements in the horizontal plane in steps of 15°. The results revealed that directivity varies significantly for different articulations, such as between voiceless fricatives. In this context, the authors found that the directivity of an [s] is more strongly directed to the front than that of an [f].

Whereas most studies analyzed the directivity in third-octave or octave bands, recent work by Blandin *et al.* (2016) and Blandin *et al.* (2018) examined the spectral fine-structure of voice directivity based on simulations and measurements with a mouth replica. The investigations revealed that directivity could change significantly within small frequency intervals, e.g., within a range of 100 Hz. Most recently, Brandner *et al.* (2020) compared such simulations to measurements and examined the influence of the mouth opening on directivity. The results showed that change in voice directivity while singing can be explained by variations of the mouth opening size.

Only a few of the studies mentioned above are based on full-spherical measurements of the human voice directivity. Moreover, the measurements were mostly carried out sequentially, i.e., separately for each direction. Thus, the spectrum of the radiated sound can only be analyzed averaged over time, and time-variant effects, often referred to as dynamic directivity, cannot be considered. Accordingly, to analyze this specific characteristic of human voice radiation, the measurements need to be carried out simultaneously for all measurement positions, e.g., by using a surrounding spherical microphone array (Arend *et al.*, 2017; Arend *et al.*, 2019; Pollow, 2015), which allows determination of the sound radiation in all directions. Up to now, only a limited number of studies based on spherical measurements have been published, e.g., Kob and Jers (1999), Chu and Warnock (2002), and Leishman *et al.* (2021), but to the best of our knowledge, there is no publicly available database. Kob and Jers (1999) focused on developing an artificial singer, and thus the study lacks detailed values on voice directivity. Chu and Warnock (2002) provided detailed values, including mean and standard deviations (SDs) of measurements for 40 subjects and a Brüel & Kjær (Nærum, Denmark) head and torso simulator in third-octave bands for 92 positions along a spherical grid. Recently, Leishman *et al.* (2021) published results of measurements with six subjects, carried out with a resolution of 5° in the horizontal and vertical planes. However, as both Chu and Warnock (2002) and Leishman *et al.* (2021) averaged over a complete sentence of fluent speech, the data do not provide any information on time-variant effects or articulation dependencies.

For research and also for applications, a spatially continuous representation of the directivity is desired. Ideally, measurements on a dense grid that can be interpolated without significant artifacts can be obtained, as shown, for example, by Leishman *et al.* (2021). However, measurements of dynamic directivity with a surrounding spherical microphone array are usually spatially sparse. To obtain a spatially continuous description of the dynamic directivity, they need to be upsampled with a specific interpolation method that minimizes so-called sparsity errors as efficiently as possible. For this purpose, several methods have been proposed, e.g., by Pollow (2015) or Ahrens and Bilbao (2021). Recently, in Pörschmann and Arend (2020a), we presented a method for spatial upsampling of sparse directivity datasets in the spherical harmonics (SH) domain [Williams (1999), Chap. 6; Rafaely (2015), Chap. 1). The so-called spatial upsampling by directional equalization (SUpDEq) method, which we originally proposed for spatial upsampling of sparse head-related transfer function (HRTF) measurements (Pörschmann *et al.*, 2019), reduces the spatial complexity of a sparse spherical directivity set by directional equalization of the measurements before the SH transform and interpolation. To evaluate the method, we applied it to a measured directivity pattern of a HEAD Acoustics (Herzogenrath, Germany) HMS II.3 head and mouth simulator and showed that the approach significantly reduces sparsity errors and thus allows us to determine a meaningful high-resolution voice directivity from sparse measurements (Pörschmann and Arend, 2019a, 2020a). In particular, we showed that for spherical directivity measurements obtained with a 32-channel array, the average spectral deviations compared with a reference measurement on a dense grid with 2702 directions are below 4 dB at frequencies up to 8 kHz. Whereas the errors are comparably large for rearward sound radiation, they are much lower in the frontal region. Finally, we discussed in Pörschmann and Arend (2020a) how the proposed method could be applied to human voice directivity measurements instead of a dummy head and mouth simulator. In Pörschmann and Arend (2020b), we presented a first approach to apply this method to human voice articulations.

This study extends the state of research in various ways. We present dense spherical datasets, simultaneously measured for all sampling positions for numerous phonemes and subjects. To our best knowledge, no comparable database of phoneme-dependent spherical voice directivities has been published before. The presented measured and spatially upsampled datasets allow the determination of directivity patterns in the horizontal and vertical planes, analysis of the spherical directivities, comparison of different phonemes, and investigation of inter-subject variances. The datasets, as well as insights from the analysis, could then be used to model human voice production. Furthermore, certain measures, such as directivity indices (DIs), can be calculated based on the datasets, and statistically significant differences in human voice radiation of different phonemes can be elaborated. Finally, due to the spatial density, the datasets can be included in auralizations of the human voice in virtual acoustic environments. They are also well-suited for visualization, be it spatial, frequency-dependent, or a combination of both. Based on these datasets, we consolidate and confirm

https://doi.org/10.1121/10.0005401

numerous investigations of other research groups, which were often hard to compare because they were carried out using different measurement procedures and datasets.

The paper is structured as follows. Section II describes the measurement procedure to acquire the directivity datasets used in this study. Section III analyzes the directivity measurements depending on articulation and the individual variances. Section IV discusses the results of this analysis, and Sec. V concludes the article and provides an outlook on how the measured directivity data can be applied for auralizing human actors in virtual reality (VR) or augmented reality (AR).

II. MATERIALS

A. Measurements

We performed the measurements in the anechoic chamber of TH Köln, which has a size of 4.5 m × 11.7 m × 2.3 m (width × depth × height) and a lower cut-off frequency of about 200 Hz. For the simultaneous measurement, we used a surrounding spherical microphone array (Arend *et al.*, 2017; Arend *et al.*, 2019), with a diameter of 2 m and a shape of a pentakis dodecahedron with 32 cardioid microphones [RØDE Microphones (Silverwater, Australia) NT5] located at the vertices. This sampling scheme allows resolving the directivity up to a SH order of $N = 4$ (Pollow, 2015). An additional microphone of the same type was positioned at the front at an azimuth of $\phi = 0°$ and an elevation of $\theta = 0°$ and served as reference. Four RME Audio (Haimhausen, Germany) Octamic II devices were used as preamplifiers and analog-to-digital/digital-to-analog (AD/DA) converters for the 32 microphones of the array. All signals were managed with two RME Fireface UFX audio interfaces. One of these audio interfaces was also used as a preamplifier and AD/DA converter for the reference microphone. For a more detailed description of the microphone array setup, please refer to Arend *et al.* (2017) and Arend *et al.* (2019).

To place the subject's head precisely in the center of the array, we adapted the seat's height and used a cross-line laser to adjust the head in all three dimensions. During the measurements, the subject's position was continuously monitored by the operator, and if required, instructions for a readjustment were given. Figure 1 shows the described measurement setup with a person sitting inside the surrounding spherical microphone array. In addition to some of the microphones of the array, the picture also shows the reference microphone and one of the cross-line lasers (left outside of the array).

We measured articulations of five vowels ([a], [e], [i], [o], [u]) and three fricatives ([f], [s], [ʃ]) two times each for 13 subjects (2 females and 11 males) aged between 25 and 64 years. None of the subjects sang professionally, but some of them did so as amateurs. The vowels were measured using the glissando method, i.e., the subjects sang a vowel with an increasing pitch over at least an octave, as proposed by Kob and Jers (1999). To measure the fricatives, the subjects articulated the respective consonant for at least 3 s.

FIG. 1. (Color online) Person inside the surrounding spherical microphone array during a measurement.

To obtain the optimal head radius required to calculate a subject-specific spherical head model in the spatial upsampling process (see Sec. II C), we determined the head width [mean (M) = 15.1 cm; SD = 0.76 cm], the head height (M = 12.6 cm; SD = 1.35 cm), and the head length (M = 19.5 cm; SD = 0.86 cm) for each subject.

B. Postprocessing

The postprocessing and spatial upsampling outlined in this section and Sec. II C were carried out in the same way as described in greater detail in Pörschmann and Arend (2020a), utilizing the SUpDEq toolbox (Audio Group Cologne, 2019; Pörschmann *et al.*, 2019). In general, the directivity analysis can be based either on time frames of the voice signals or on impulse responses determined for the articulation of different phonemes, which is the approach we took in this study. Thus, as a first step, impulse responses between each signal of the array microphones and the reference microphone were calculated for each of the articulated phonemes.

The directivity measurements might be influenced toward low frequencies by reflections and room modes of the anechoic chamber, which in our case becomes relevant below 200 Hz and thus affects the measurements around the human fundamental frequency. Therefore, we applied a low-frequency extension according to Xie (2009), substituting the original low-frequency component in the frequency domain by an adequately matched component. Finally, the impulse responses were windowed and truncated to a length of 128 samples at a sampling rate of 48 kHz, which removes (potential) reflections at the boundaries of the microphone array and reduces the computational power required when using these filters for auralizations in virtual acoustic environments (VAEs).

As discussed in detail in Pörschmann and Arend (2020a), another postprocessing step is required to compensate for inaccuracies when positioning the subject in the center of the array. Furthermore, the setup of the pentakis dodecahedron implies minor inaccuracies of the microphone

positions. Both result in slightly varying distances from the center of the head to the microphones. Our analysis in Pörschmann and Arend (2019b) revealed that even slight deviations lead to substantial impairments in the spatial upsampling. Several approaches have been described to overcome these influences of distance inaccuracies, e.g., by Ben Hagai *et al.* (2011) or Richter *et al.* (2014). In this study, we applied a method for distance error compensation that successfully reduced the impairments of distance errors for voice directivities of a dummy head with an integrated mouth simulator (Pörschmann and Arend, 2020a).

C. Spatial upsampling

In the next step, we applied the SUpDEq method (Pörschmann *et al.*, 2019) to spatially upsample the sparse datasets with 32 sampling points, as obtained from postprocessing, to a dense grid with 2702 sampling points on a Lebedev sampling scheme. As a comprehensive description is far beyond the scope of the present work, the reader is referred to Pörschmann and Arend (2020a), where we elaborated and evaluated the SUpDEq method for spatial upsampling of voice directivity measurements in detail. In the following, we briefly describe the basic idea.

In the equalization, the datasets are modified so the mouth is virtually shifted to the center of the head, resulting in a time-alignment (and spectral alignment) of the dataset. As most energy in higher spatial orders of non-aligned datasets originates from rapid phase changes between neighboring directions caused by the off-center location of the mouth, the equalized time-aligned datasets show significantly lower energy in higher spatial orders and thus have lower spatial complexity. Accordingly, errors due to spatial aliasing and order truncation are significantly decreased when equalizing the dataset before the transformation to the SH domain. The equalization is carried out by spectral division with corresponding directional rigid sphere transfer functions. The directional rigid sphere transfer functions describe the sound radiation from a point on the rigid sphere's surface into the far-field and can be analytically determined for any radiation direction. They represent a simplified voice directivity carrying no information on the specific shape of the mouth opening or the form of, for example, the cheekbones, but only featuring the basic shape of a spherical head.

Next, spatial upsampling is performed by applying an inverse SH transform on a dense grid, resulting in a dense equalized dataset with interpolated values. Finally, the directivity pattern is de-equalized by a spectral multiplication with corresponding directional rigid sphere transfer functions, which inverts the alignments and recovers the temporal and spectral properties. As analyzed in detail in Pörschmann and Arend (2020a), for a dummy head with mouth simulator, the SUpDEq method allows determination of directivity patterns that are much closer to the reference than directivity patterns obtained with common SH interpolation without any pre- and postprocessing.

For spatial upsampling, it is advantageous to adjust the radius of the spherical head model used for equalization and de-equalization according to each subject. In the present case, the optimal radius for each subject was calculated according to Algazi *et al.* (2001) based on measurements of the head width, height, and length. As the main voice radiation direction, we defined $\phi = 0°$ and $\theta = -25°$, which is in line with Marshall and Meyer (1985). A more detailed analysis of the SUpDEq method's performance and how the inaccuracies depend on radiation direction can be found in Pörschmann and Arend (2020a).

III. RESULTS

A. Horizontal and vertical planes

To analyze the voice directivity for the different phonemes on the horizontal and vertical planes, the postprocessed and spatially upsampled datasets were transformed to the SH domain at a SH order $N = 35$ and spatially resampled to 360 directions in 1° steps along the horizontal plane ($\phi = -180°$ to 180°, where positive angles point to the left; $\theta = 0°$) and along the vertical plane ($\phi = 0°; \theta = -180°$ to 180°, where 0° points to the front, 90° to the top, and 180° to the back) using the inverse SH transform. We averaged the resulting directivity patterns over the two repeated measurements for each subject and phoneme.

Figures 2 and 3 show the directivity patterns for the vowels and the fricatives in octave bands with center frequencies from 250 Hz to 8 kHz. To illustrate interindividual variations between the subjects, each plot shows mean values as well as the SDs for one phoneme. We refrained from showing the directivity patterns for frequencies below 250 Hz because no differences in the directivity can be observed for the low frequencies between the different phonemes, and we applied the low-frequency extension below 200 Hz. Toward higher frequencies, we only show plots up to the 8 kHz octave band for two reasons: First, most of the energy of human voice articulation is far below 8 kHz, even though some recent studies claim that higher frequencies also carry important information (Blandin *et al.*, 2018; Kocon and Monson, 2018). Second, as shown in Pörschmann and Arend (2020a), the errors caused by upsampling the datasets measured on the sparse sampling grid can become quite large for frequencies above 8 kHz.

The plots in Figs. 2 and 3 clearly show that below 1 kHz, the directivity patterns vary only slightly between the different articulations and that the directivity becomes stronger directed to the front with increasing frequency. In general, differences between the phonemes increase with frequency as well. For example, in the horizontal plane, there are significant differences between an [a] and the other vowels in the octave bands of 4 and 8 kHz, and the radiation is stronger directed to the front for an [a]. However, these differences seem to diminish in the vertical plane, even toward higher frequencies. Furthermore, with increasing frequency, specific peaks and dips can be observed in the rearward area, probably caused by constructive and destructive

https://doi.org/10.1121/10.0005401

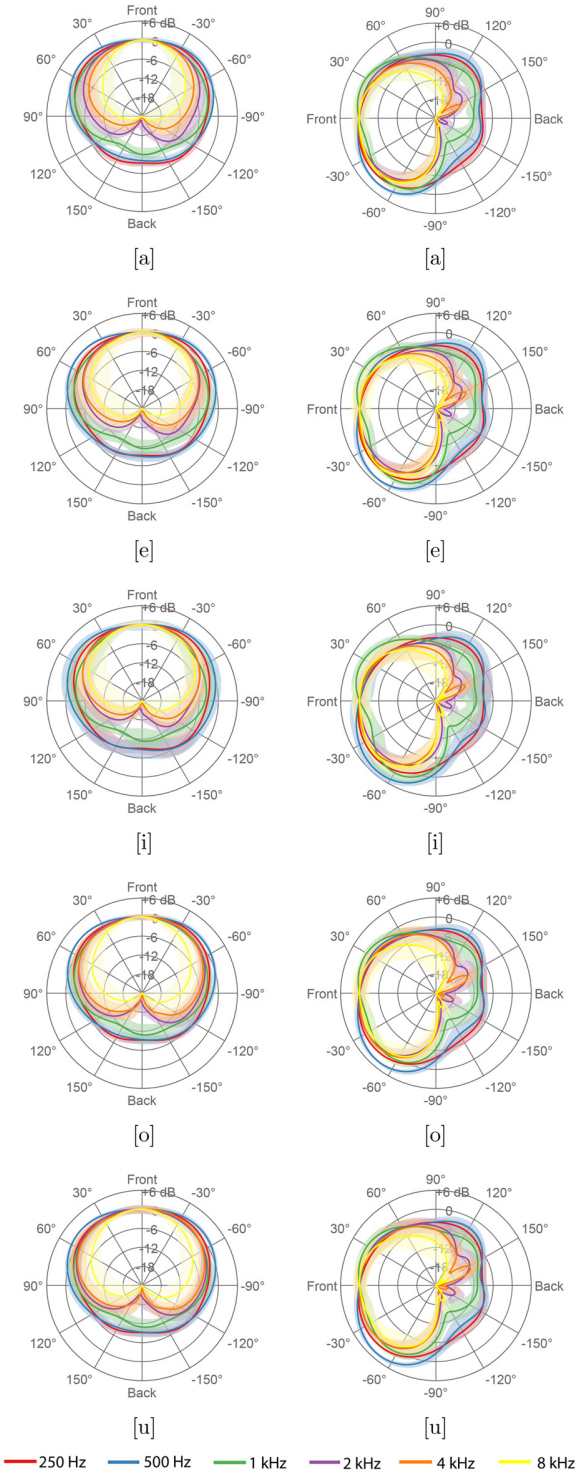


FIG. 2. (Color online) Directivity in the horizontal plane (left column) and the vertical plane (right column) for the vowels [a], [e], [i], [o], [u]. Shown are the mean values and the SDs of the interindividual variations in the octave bands with center frequencies between 250 Hz and 8 kHz.

interferences of the sound diffracted around the head. This can be observed quite well for the plots in the vertical plane. In the frontal hemisphere, for frequencies up to 2 kHz, the SDs do not exceed 2 dB for the vowels and 3 dB for the fricatives. In general, there is a trend of increasing SDs from

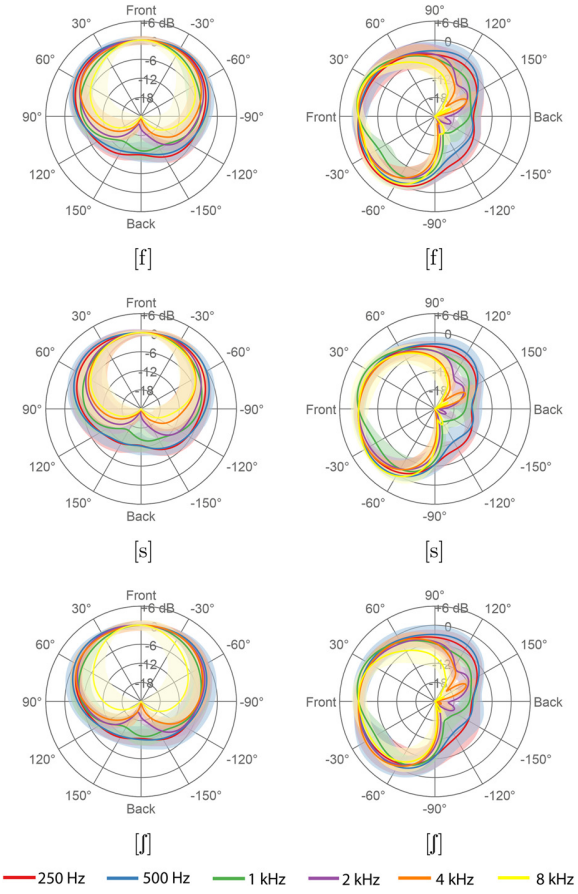


FIG. 3. (Color online) Directivity in the horizontal plane (left column) and the vertical plane (right column) for the fricatives [f], [s], and [ʃ]. Shown are the mean values and the SDs of the interindividual variations in the octave bands with center frequencies between 250 Hz and 8 kHz.

frontal to rearward sound radiation. For lateral and rearward sound radiation, the SDs tend to be larger for the fricatives than for the vowels, especially in the octave bands of 4 and 8 kHz. For example, for lateral sound radiation ($\phi = \pm 90^\circ$) at 4 kHz, the SD reaches about 4 dB for the vowels and 6 dB for the fricatives. However, this increase at rearward directions might, to some extent, be influenced by the spatial upsampling, which, as shown in Pörschmann and Arend (2020a), causes maximal deviations for high frequencies at rearward directions.

As shown in the polar plots in Figs. 2 and 3, differences between the phonemes are hard to discriminate. Thus, in the following, we analyze the directivity patterns based on spectral differences. However, as there is no reference measurement for comparison, we calculated for each phoneme the spectral deviations ΔG_{sp} to a directivity pattern obtained by averaging over all articulations. Analyzing ΔG_{sp} allows identifying frequency ranges or spatial regions in which the directivity is increased or reduced by a specific articulation. For this, we first equalized each measured dataset to a normalized radiation power averaged over all directions. Then we determined the directivity pattern D_{ph} per phoneme averaged over all subjects and the directivity pattern D_{av}

averaged over all articulations and subjects. The spectral deviations per direction were then calculated in dB as

$$\Delta G_{sp}(\theta,\phi,f) = 20lg\frac{|D_{ph}(\theta,\phi,f)|}{|D_{av}(\theta,\phi,f)|}. \quad (1)$$

We determined ΔG_{sp} for each articulation on a full-spherical test sampling grid generated for ϕ and θ in steps of 1° and averaged the values over all subjects.

Figures 4 and 5 show ΔG_{sp} in the horizontal and vertical planes for the vowels and the fricatives, respectively. The plots indicate that in the horizontal and vertical planes, the frontal sound radiation of an [a] is stronger than for the other phonemes, especially for frequencies above 2 kHz. Furthermore, for an [e] and an [i], the plots show an increase in the radiation directed downward, mainly in the frequency range up to 2 kHz, and for a [u] between 1.5 and 6 kHz. For [o] and [u], the radiation is less directed to the front and

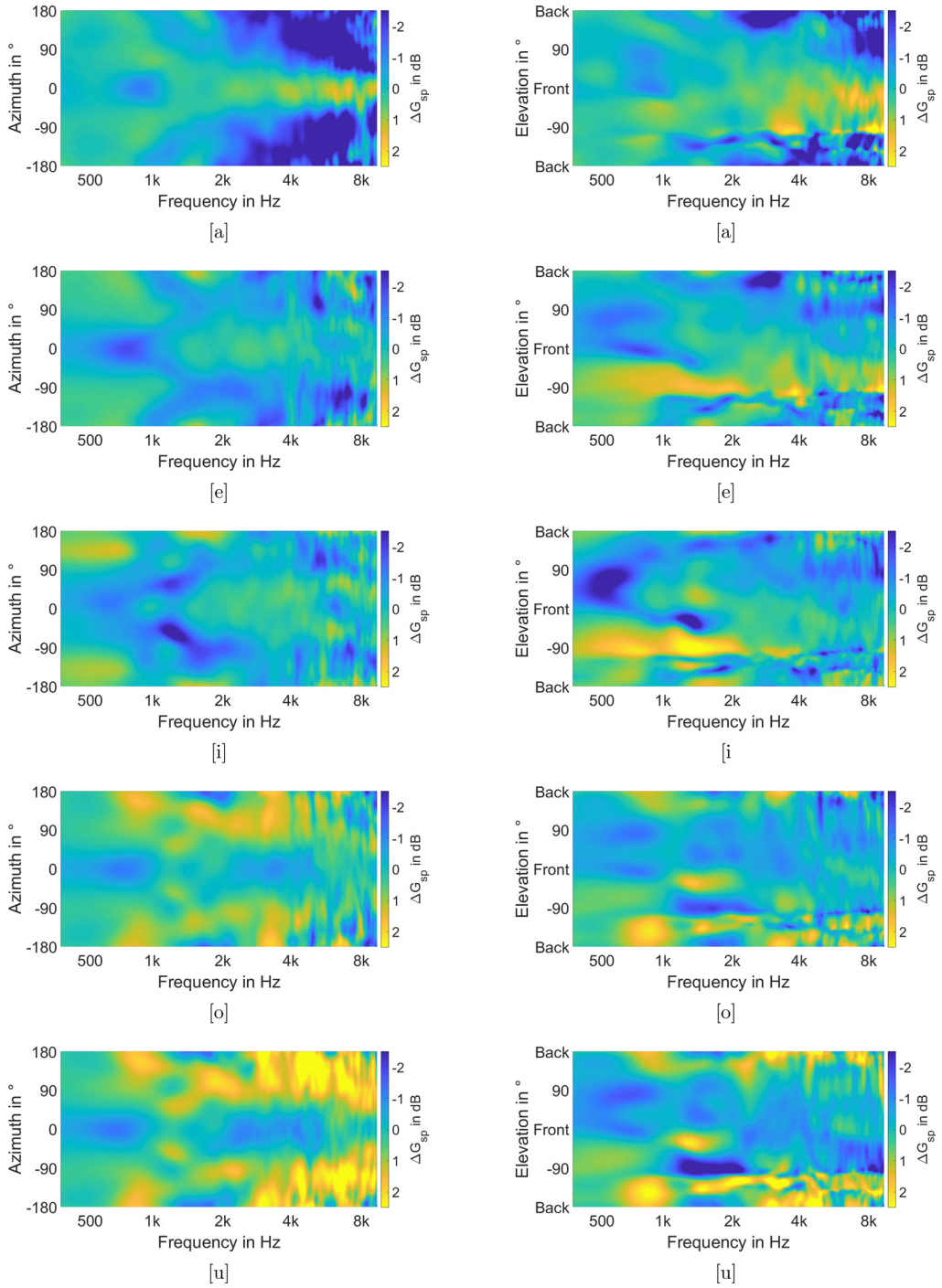


FIG. 4. (Color online) Spectral deviations ΔG_{sp} between the vowels and the mean of all phonemes. Left column, horizontal plane; right column, vertical plane.

https://doi.org/10.1121/10.0005401

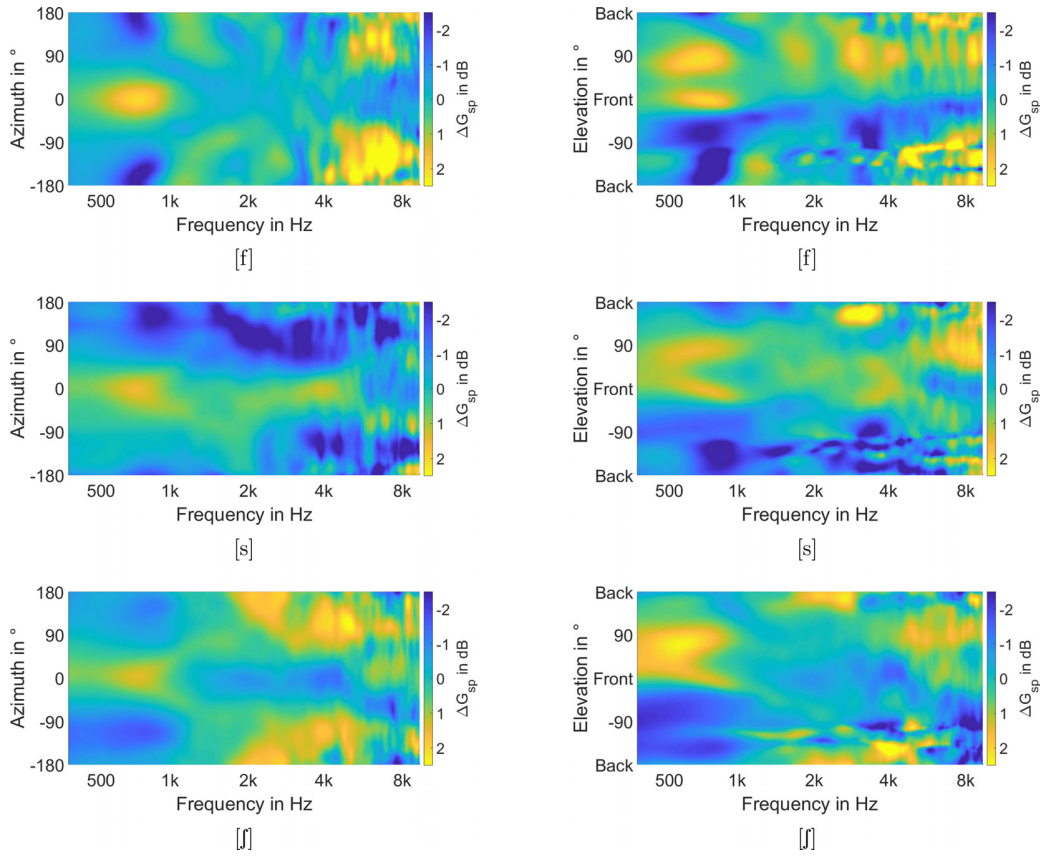


FIG. 5. (Color online) Spectral deviations ΔG_{sp} between the fricatives and the mean of all phonemes. Left column, horizontal plane; right column, vertical plane.

increases in the rear hemisphere compared to the average values. All these effects seem to be stronger for a [u] than for an [o]. All fricatives tend to have increased sound radiation directed upward, which is strongest in the frequency range between 500 Hz and 1 kHz. Furthermore, we observed an increased frontal radiation above 1 kHz for an [f] and an [ʃ].

B. Full-spherical

This section focuses on the analysis of the spherical datasets considering the deviations between the different articulated phonemes. Figures 6 and 7 show the spectral deviations ΔG_{sp} for the articulation of vowels and fricatives averaged over frequency for $f \leq 8$ kHz. For the vowels, Fig. 6 indicates similarities for an [a], [e], and [i], which show a slightly increased frontal radiation that is strongest and most focused for an [a]. For [o] and [u], an increased sound radiation around $\phi = \pm 120°$, $\theta = 0°$ can be observed, which is stronger for a [u]. Regarding the fricatives, Fig. 7 indicates regions with a moderately increased sound radiation at about $\phi = \pm 90°$, $\theta = 45°$ for [f] and [ʃ]. Furthermore, this increase seems to be slightly asymmetric for [ʃ] with higher values for sound radiation to the left. Finally, as already analyzed based on Figs. 4 and 5, the radiation is directed slightly downward for [a], [e], and [i] and upward for the fricatives.

C. Directivity index

As a measure describing directivity by a single frequency-dependent value, we determined the DI for spherical sound radiation according to Long (2014), which can be calculated as

$$DI(f) = 10\,lg \frac{4\pi |D(\theta_0, \phi_0, f)|^2}{\int_0^{2\pi} \int_{-\pi/2}^{\pi/2} |D(\theta, \phi, f)|^2 \cos\theta d\theta d\phi}, \quad (2)$$

with ϕ_0, θ_0 the frontal direction.

To get a better general understanding of how much the DIs are affected by spatial upsampling errors, we analyzed in an informal evaluation measurements from our previous study (Pörschmann and Arend, 2020a) and compared DIs of dense reference datasets to DIs of upsampled datasets from measurements with the surrounding spherical microphone array. Averaged over frequencies up to 8 kHz, the DI (third-octave smoothed) varied less than 0.4 dB, with a maximum value of 1.3 dB at 2.25 kHz. We assume that the errors due to upsampling are so small that significant changes in DI are caused by the different phonemes.

Through statistical analysis, we examined the DIs for the 13 subjects in more detail concerning their frequency- and phoneme-dependent differences. The analysis on the

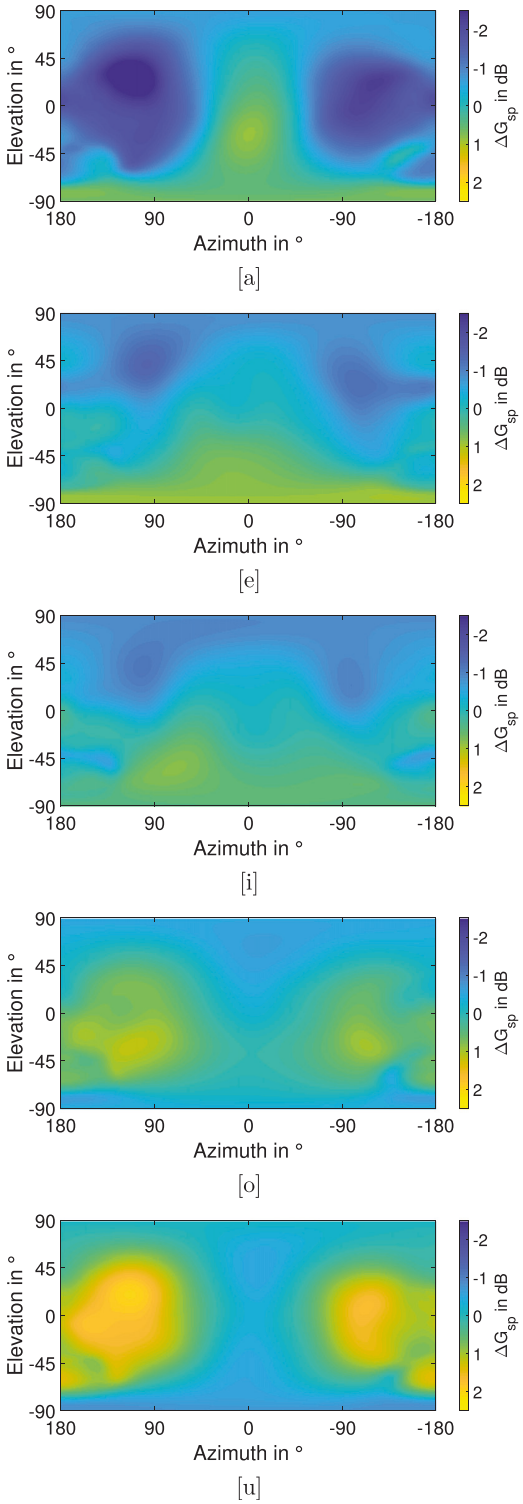


FIG. 6. (Color online) Spectral deviations ΔG_{sp} between the vowels and the mean of all phonemes for $f \leq 8$ kHz.

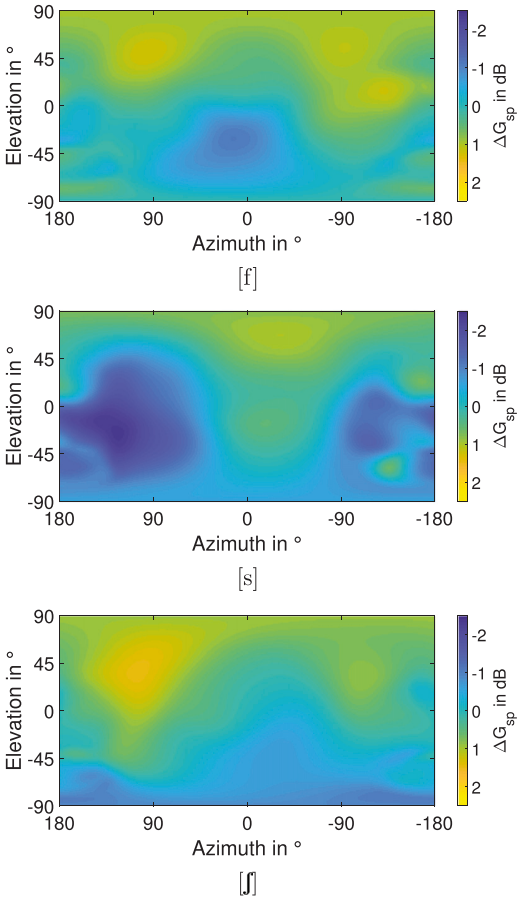


FIG. 7. (Color online) Spectral deviations ΔG_{sp} between the fricatives and the mean of all phonemes for $f \leq 8$ kHz.

phoneme-dependencies was based on the mean value per subject, i.e., the average of the two DIs per condition determined from the two measurements per subject. Shapiro–Wilk tests for normality with Hochberg correction (Hochberg, 1988) for multiple hypothesis testing yielded no violation of normality for any of the 152 tested conditions (19 third-octave frequency bands from 125 Hz to 8 kHz, eight phonemes). Thus, we analyzed the DIs with a two-way repeated measures analysis of variance (ANOVA) with the within-subject factors frequency and phoneme, nevertheless corrected for slight violations of ANOVA assumptions using the Greenhouse–Geisser (GG) correction (Greenhouse and Geisser, 1959). The ANOVA yielded a significant main effect of frequency [$F(18,216) = 63.20$, $p_{GG} < 0.001$, $\eta_p^2 = 0.84$, $\epsilon = 0.17$] and phoneme [$F(7,84) = 27.33$, $p_{GG} < 0.001$, $\eta_p^2 = 0.69$, $\epsilon = 0.52$] as well as a significant frequency × phoneme interaction [$F(126,1512) = 9.36$, $p_{GG} < 0.001$, $\eta_p^2 = 0.44$, $\epsilon = 0.06$].

Figure 8 shows the mean values and the SDs of the DI for the different phonemes, clearly illustrating the significant influence of frequency and phoneme. The differences between vowels and fricatives are evident, especially for frequencies below 1 kHz. For both vowels and fricatives, there is a minimum of the DI at a phoneme-dependent frequency between 500 and 700 Hz. Whereas for the vowels, the minima are below the phoneme-average (black curve), for the fricatives, they are about 2 dB higher and above the average. Furthermore, the shape of the minima is less prominent for the fricatives. A more detailed comparison reveals

https://doi.org/10.1121/10.0005401

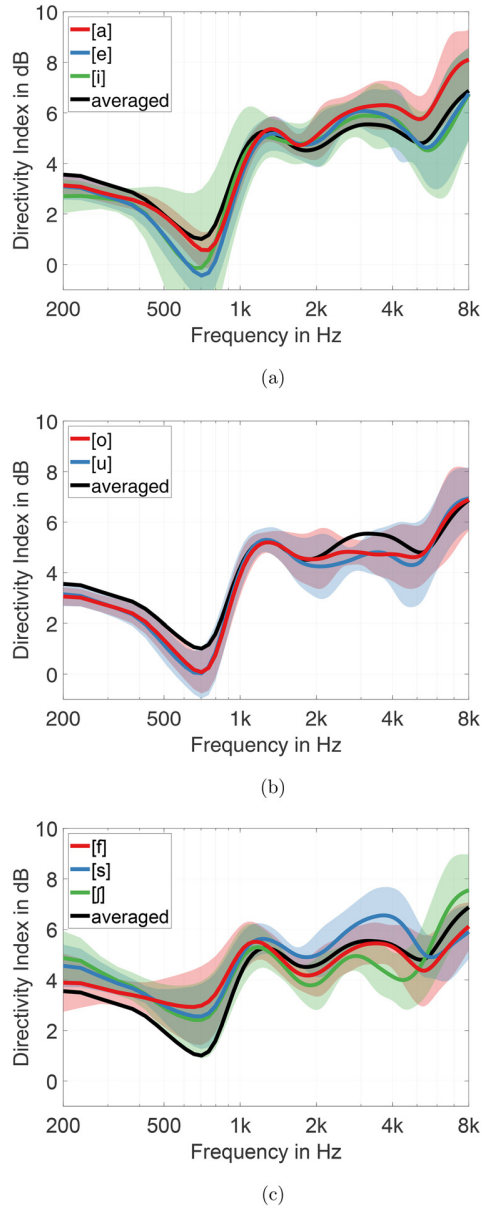


FIG. 8. (Color online) Mean values and SDs of the DI averaged over all subjects for the different articulations (third-octave smoothed) related to the frontal direction. (a) [a], [e], [i]; (b) [o], [u]; (c) [f], [s], [ʃ]. The black curve shows the average over all phonemes.

that, in the frequency range between 2 and 4 kHz, the curves for [a], [e], and [i] are above the average curve, whereas for [o] and [u], they are below the average curve. The SDs are in the frequency range between 400 Hz and 1.5 kHz for an [i] larger than for the other phonemes, reaching a maximum of more than ±3 dB. In contrast, the SDs are below ±1 dB for all other vowels.

To further support these observations on phoneme-dependent differences, we performed more detailed statistical analysis. In this sense, we averaged the DIs over the five vowels or the three fricatives, resulting in frequency-dependent averaged DIs for the two phoneme types vowels and fricatives. By comparing the DIs concerning phoneme type, we examined whether there is a significant difference between the frequency-dependent mean of all vowels and fricatives. A GG-corrected two-way repeated measures ANOVA with the within-subject factors frequency and phoneme type revealed a significant main effect for phoneme type [$F(1,12) = 56.34$, $p_{GG} < 0.001$, $\eta_p^2 = 0.86$, $\epsilon = 1$], which statistically confirms the distinct difference in DIs for vowels and fricatives. For this reason, in the further statistical analysis, we examined separately for both phoneme types (vowels and fricatives) to what extent individual phonemes result in significantly different DIs and in which frequency bands significant differences in the DIs occur.

To examine which vowels and fricatives lead to significantly different DIs, we performed pairwise comparisons between all vowels (five vowels, resulting in ten pairwise comparisons) and all fricatives (three fricatives, resulting in three pairwise comparisons) using nested GG-corrected two-way repeated measures ANOVAs with the within-subject factors frequency and articulation. The initial significance level 0.05 was further corrected according to Hochberg to prevent alpha-error accumulation. For the vowels, the ANOVAs yielded no significant main or interaction effect of articulation for the four pairwise comparisons [e] vs [i], [i] vs [o], [i] vs [u], and [o] vs [u] (all main effect $p_{GG} > 0.342$, all interaction effect $p_{GG} > 0.052$; for the sake of conciseness, the statistical results of the nested ANOVAs are not reported in greater detail throughout the paper). The statistical results suggest similar frequency-dependent DIs for mentioned vowels, which can be seen in Fig. 8, especially when comparing [e] vs [i] or [o] vs [u]. Overall, the statistical results indicate that, in particular, the DIs for an [a] differ significantly compared to the other vowels, whereas especially [i], [o], and [u] lead to very similar frequency-dependent DIs. For the fricatives, the ANOVAs revealed significant main effects of articulation for [f] vs [s] ($p_{GG} = 0.020$) and [f] vs [ʃ] ($p_{GG} = 0.004$) as well as significant frequency × articulation interaction effects for all three pairwise comparisons (all interaction effect $p_{GG} < 0.008$). Thus, the statistical results revealed distinct frequency-dependent differences between the DIs for the fricatives, thereby confirming the observations made above based on Fig. 8.

Finally, we analyzed, separately for the vowels and fricatives, in which frequency bands the DIs for the different phonemes vary significantly. For this, we conducted a nested GG-corrected one-way repeated measures ANOVA with the within-subject factor phoneme (five levels for vowels, three for fricatives) for each of the 19 levels of frequency. Again, Hochberg correction was applied to correct for multiple hypothesis testing. For the vowels, the ANOVAs revealed significant differences in DIs for the third-octave bands centered at 315 Hz, 2 kHz, 2.5 kHz, 3.15 kHz, 4 kHz, and 6.3 kHz (all $p_{GG} < 0.003$). However, especially below 1.5 kHz, the DIs for an [i] show significantly higher inter-subject variance than the other vowels (see Fig. 8), so we performed the same analysis again, excluding the DIs for [i] (i.e., four-way repeated measures ANOVAs for each of the

19 levels of frequency). These ANOVAs yielded significant differences in DIs for all third-octave bands from 200 Hz to 1 kHz and, consistent with the previous analysis including the DIs for an [i], significant differences in DIs for the third-octave bands from 2 to 6.3 kHz (all $p_{GG} < 0.002$). The analysis indicates that the high inter-subject variance of the DIs for [i] caused the non-significant effect of vowel at frequencies below 1 kHz. The results excluding [i] suggest that the vowels [a], [e], [o], and [u] exhibit significant differences in DI in almost all of the analyzed third-octave bands and, thus, over a wide frequency range. For the fricatives, the ANOVAs yielded significantly different DIs for the third-octave bands centered at 1.6, 2, 2.5, 3.15, 4, 5, and 6.3 kHz (all $p_{GG} < 0.003$). Thus, within both groups, vowels and fricatives, the main differences are in the frequency range between 2 and 6 kHz.

IV. DISCUSSION

We presented spherical datasets of directivity patterns measured for 13 subjects and eight different phonemes. We analyzed the directivity patterns in the horizontal and vertical plane and studied the differences between the phonemes based on the spherical datasets. Based on the DIs, we determined statistically significant differences in human voice radiation of different phonemes.

The analysis of the directivity patterns' SDs suggests that interindividual differences tend to be maximal for rearward directions, which becomes evident when examining the SDs in Figs. 2 and 3. This can be explained by the propagation of sound around the head or a rigid sphere. Rearward, the directivity is dominated by interferences, which change rapidly for adjacent directions, especially toward higher frequencies. Even though this behavior is generally the same for all phonemes and all subjects, the detailed structure varies. Consequently, peaks and dips are not located in the same directions for the different phonemes and the different subjects. However, when plotting average datasets as in our study, peaks and dips that have been observed in other studies (e.g., Katz and D'Alessandro, 2007) are hardly present.

The results of the studies by Chu and Warnock (2002) analyzing spherical sound radiation and by Moreno and Pfretzschner (1978) analyzing the radiation in the horizontal and vertical planes match our measurements. In the frontal hemisphere, differences in the directivity patterns are mostly within the SD of our measurements. For rearward directions, some larger differences occur, which could be explained by the limited spatial resolution of the measurements in those studies not allowing the resolution of the spatially rapidly changing patterns in the rear area. The directivity patterns vary significantly between the different vowels and are strongest directed to the front for an [a], which is in line with the findings of other studies, e.g., Kocon and Monson (2018). In the same way as in our study, the phoneme-dependent differences determined by Marshall and Meyer (1985) indicated that an [e] is more directed to the front than an [o] or a [u]. Analyzing the DIs of the different vowels, we found a decrease in the following order: [a], [e], [i], [o], [u]. Considering that in our study, the differences in the DIs between [e] and [i], as well as between [i], [o], and [u], were not statistically significant, our study is in line with the results of Kocon and Monson (2018). However, in Kocon and Monson (2018), only the sound radiation in the horizontal plane was analyzed. Thus, direct comparison of the specific values for the DIs as shown in Fig. 8 is not possible. Leishman *et al.* (2021) measured DIs for spherical sound radiation averaged over a complete sentence and determined a minimum in DI in the 800 Hz third-octave band as well a substantial rise between 800 Hz and 1.6 kHz. These findings are consistent with the course of the average DI shown in Fig. 8. Moreover, the frequencies of the successive minima and maxima determined by Leishman *et al.* (2021) are very similar to those determined by us, and the DIs estimated in their study differ on average by less than 1 dB from those determined by us.

In line with Brandner *et al.* (2020), our results also showed that an increase in the mouth opening leads to a stronger directivity. Whereas the largest mouth opening occurs for an [a] and results thus in the highest DI, the lowest DIs are observed for an [o] and a [u] corresponding to the smallest mouth opening sizes. Between the fricatives, we also observed significant differences. In the frequency bands of 2 and 4 kHz, the directivity is narrower for an [s] than for an [f] or an [ʃ]. This is in line with Monson *et al.* (2012), who found that in the horizontal plane, the directivity of an [s] is more directional than that of an [f]. The results are supported by the analysis of the DIs, which showed that the differences of the DIs between [f] and [s] as well as between [f] and [ʃ] are statistically significant.

By analyzing the DIs, we showed that after averaging over all subjects, phoneme-dependent differences do not exceed 3 dB for frequencies up to 8 kHz. However, the differences in the DIs of fricatives and vowels are statistically significant. The strength of a first minimum between 500 and 700 Hz varies between the different types of phonemes, and the minimal DI is lower for vowels than for fricatives. Analyzing spatial radiation in more detail showed a further difference between vowels and fricatives, which relates to the sound radiation in the vertical plane. Whereas we observed for the vowels, especially for [a], [e], and [i], increased sound radiation slightly downward, we found for the fricatives increased radiation upward. To our knowledge, these systematic differences between fricatives and vowels have not yet been investigated in detail. Therefore, more phonemes, including some plosives, should be investigated in future studies.

When applying measured, dynamic directivity patterns for virtual acoustic environments, the perceptual influence compared to a static directivity needs to be considered. Research by Postma and Katz (2016) and Postma *et al.* (2017) indicated that auralizations involving dynamic voice directivity are perceived more plausible and exhibit a wider apparent source width than auralizations with static voice directivity or omnidirectional sources. On the contrary, in a

recent study by Ehret *et al.* (2020), the integration of dynamic, phoneme-dependent directivities was perceptually not distinguishable from a static (averaged) speaker directivity. When frontally facing a human speaker, the DI is directly related to the direct-to-reverberant energy ratio (DRR) in a room. In this context, Frank and Brandner (2019) investigated the just-noticeable differences (JNDs) of the DI for sound source reproduction based on artificially designed and frequency-independent directivity patterns. They found that for a DRR of 0 dB and frontal sound radiation, differences in DI exceeding 1.8 dB are perceptible. This is in line with JNDs found by Larsen *et al.* (2008), who determined values of about 2–3 dB in rooms with a DRR of about 0 or +10 dB and of about 6–8 dB in rooms with a DRR of −10 or +20 dB. For the DRR of 0 and+10 dB, these JNDs are in the range of the variations of the DI between the different phonemes. For the other DRRs, the variations between the different phonemes are below the JND. Accordingly, when frontally facing a human speaker, phoneme-dependent differences in the DI might be audible in a direct comparison but are probably not of major perceptual influence, e.g., regarding the plausibility of a human speaker in a virtual acoustic environment. However, the JNDs are supposed to be lower if the interlocutors do not directly look at each other but turn their heads slightly and thus radiate speech laterally, as often observed in conversations. In this case, not only the reflections and the reverberation are affected by the voice directivity, but also the direct sound component. Accordingly, for a facing angle of 30°, Frank and Brandner (2019) determined an increased sensitivity and a JND of 1 dB. However, as the DRR is a broadband measure, a prediction of how a decrease in the DI in specific frequency bands can become audible is hardly possible.

V. CONCLUSION

We determined and analyzed articulation-dependent full-spherical directivity patterns of different vowels and fricatives from two repeated measurements with 13 subjects. To obtain high-density data of human directivity patterns from sparse measurements in a surrounding spherical microphone array, we applied the SUpDEq method for spatial upsampling.

We examined five vowels and three fricatives and determined significant differences between the directivities of the phonemes. In agreement with earlier studies, we found for the vowels the strongest directivity for an [a]. Analyzing the DIs in more detail showed a tendency of decreasing directivity with decreasing mouth opening size, especially in the frequency range between 2 and 4 kHz. Furthermore, for the vowels, the analysis showed strong similarities between [i], [o], and [u] as well as for [e] and [i]. The pairwise comparisons of the other phonemes revealed significantly different DIs. Regarding the fricatives, we found significant differences between the DIs for [f] and [s] and between the DIs for [f] and [ʃ]. Generally, the statistical analysis of the DIs revealed significant differences between vowels and fricatives. Finally, based on the analysis of the spherical datasets, we found a tendency that the directivity is more directed upward for the fricatives than for the vowels.

The analysis of the spherical datasets determined in this study provides general insight into articulation-dependent aspects of human voice directivity and can contribute to models of human voice production in a more general way. In follow-up studies, these results need to be compared to directivity patterns of fluent speech. The datasets can be applied in the fields of VR and AR and in room simulations to integrate adequate radiation patterns in the process of sound-field calculation and rendering. In this context, we plan to perceptually evaluate how accurately voice directivity patterns need to be reconstructed in a VAE and to what extent interindividual differences, as well as articulation-dependent variations, have to be considered. A suitable system design and procedure for rendering human voice directivity in VR and AR can be proposed based on these results.

ACKNOWLEDGMENTS

To foster reproducible research and open science, we provide a database with anonymized datasets of human voice directivities that can be used for further studies (Pörschmann and Arend, 2021). The database contains sparse measurements at the sampling positions of the pentakis dodecahedron, datasets upsampled to a dense grid of 2702 positions, and a MATLAB script to determine the upsampled from the sparse datasets and to create some basic plots. The authors would like to thank Raphaël Gillioz for supporting the measurements. The research presented in this paper has been carried out in the Research Project NarDasS, which was funded by the Federal Ministry of Education and Research in Germany, support code: BMBF 03FH014IX5-NarDasS.

Ahrens, J., and Bilbao, S. (**2021**). "Computation of spherical harmonic representations of source directivity based on the finite-distance signature," IEEE/ACM Trans. Audio Speech Language Process. **29**, 83–92.
Algazi, V. R., Duda, R. O., and Avendano, C. (**2001**). "Estimation of a spherical-head model from anthropometry," J. Audio Eng. Soc. **49**(6), 472–479.
Arend, J. M., Lübeck, T., and Pörschmann, C. (**2019**). "A reactive virtual acoustic environment for interactive immersive audio," in *Proceedings of the 2019 AES Conference on Immersive and Interactive Audio*, March 27–29, York, UK.
Arend, J. M., Stade, P., and Pörschmann, C. (**2017**). "Binaural reproduction of self-generated sound in virtual acoustic environments," Proc. Mtgs. Acoust. **30**, 015007.
Arend, J. M., and Pörschmann, C. (**2019**). "SUpDEq—Spatial upsampling by directional equalization," https://github.com/AudioGroupCologne/SUpDEq (Last viewed June 18, 2021).
Ben Hagai, I., Pollow, M., Vorländer, M., and Rafaely, B. (**2011**). "Acoustic centering of sources measured by surrounding spherical microphone arrays," J. Acoust. Soc. Am. **130**(4), 2003–2015.
Blandin, R., Hirtum, A. V., and Laboissière, R. (**2016**). "Influence of higher order acoustical propagation modes on variable section waveguide directivity: Application to vowel [A]," Acta Acust. united Acust. **102**, 918–929.
Blandin, R., Hirtum, A. V., Pelorson, X., and Laboissière, R. (**2018**). "The effect on vowel directivity patterns of higher order propagation modes," J. Sound Vib. **432**, 621–632.
Brandner, M., Blandin, R., Frank, M., and Sontacchi, A. (**2020**). "A pilot study on the influence of mouth configuration and torso on singing voice

directivity A pilot study on the influence of mouth configuration and torso on singing voice directivity," J. Acoust. Soc. Am. **148**(3), 1169–1180.
Cabrera, D., Davis, P. J., and Connolly, A. (**2011**). "Long-term horizontal vocal directivity of opera singers: Effects of singing projection and acoustic environment," J. Voice **25**(6), e291–e303.
Chu, W. T., and Warnock, A. C. C. (**2002**). Detailed Directivity of Sound Fields Around Human Talkers, Technical Report (National Research Council Canada, Ottawa, Canada).
Dunn, H. K., and Farnsworth, D. W. (**1939**). "Exploration of pressure field around the human head during speech," J. Acoust. Soc. Am. **10**, 184–199.
Ehret, J., Stienen, J., Brozdowski, C., Bönsch, A., Mittelberg, I., and Vorländer, M. (**2020**). "Evaluating the influence of phoneme-dependent dynamic speaker directivity of embodied conversational agents' speech," in *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, October 20–22.
Frank, M., and Brandner, M. (**2019**). "Perceptual evaluation of spatial resolution in directivity patterns," in *Proceedings of the 45th DAGA Congress*, March 18–21, Rostock, Germany, pp. 74–77.
Greenhouse, S. W., and Geisser, S. (**1959**). "On methods in the analysis of profile data," Psychometrika **24**(2), 95–112.
Henry, J. (**1857**). *Annual Report of the Board of Regents of the Smithsonian Institution* (Smithsonian Institution, Washington, DC).
Hochberg, Y. (**1988**). "A sharper Bonferroni procedure for multiple tests of significance," Biometrika **75**(4), 800–802.
Katz, B., and D'Alessandro, C. (**2007**). "Directivity measurements of the singing voice," in *Proceedings of the 19th International Congress on Acoustics*, September 2–7, Madrid, Spain.
Kob, M., and Jers, H. (**1999**). "Directivity measurement of a singer," J. Acoust. Soc. Am. **105**, 1003.
Kocon, P., and Monson, B. B. (**2018**). "Horizontal directivity patterns differ between vowels extracted from running speech," J. Acoust. Soc. Am. **144**(1), EL7–EL12.
Larsen, E., Iyer, N., Lansing, C. R., and Feng, A. S. (**2008**). "On the minimum audible difference in direct-to-reverberant energy ratio," J. Acoust. Soc. Am. **124**(1), 450–461.
Leishman, T. W., Bellows, S. D., Pincock, C. M., and Whiting, J. K. (**2021**). "High-resolution spherical directivity of live speech from a multiple-capture transfer function method," J. Acoust. Soc. Am. **149**(3), 1507–1523.
Long, M. (**2014**). *Architectural Acoustics* (Elsevier, Amsterdam).
Marshall, A. H., and Meyer, J. (**1985**). "The directivity and auditory impressions of singers," Acustica **58**, 130–140.
Monson, B. B., Hunter, E. J., and Story, B. H. (**2012**). "Horizontal directivity of low- and high-frequency energy in speech and singing," J. Acoust. Soc. Am. **132**(1), 433–441.
Moreno, A., and Pfretzschner, J. (**1978**). "Human head directivity in speech emission: A new approach," Acoust. Lett. **1**, 78–84.
Pollow, M. (**2015**). *Directivity Patterns for Room Acoustical Measurements and Simulations* (Logos Verlag, Berlin), pp. 1–176.
Pörschmann, C., and Arend, J. M. (**2019a**). "A method for spatial upsampling of directivity patterns of human speakers by directional equalization," in *Proceedings of the 45th DAGA Congress*, March 18–21, Rostock, Germany, pp. 1458–1461.
Pörschmann, C., and Arend, J. M. (**2019b**). "How positioning inaccuracies influence the spatial upsampling of sparse head-related transfer function sets," in *Proceedings of the International Conference on Spatial Audio—ICSA 2019*, September 26–28, Ilmenau, Germany, pp. 1–8.
Pörschmann, C., and Arend, J. M. (**2020a**). "A method for spatial upsampling of voice directivity by directional equalization," J. Audio Eng. Soc. **68**(9), 649–663.
Pörschmann, C., and Arend, J. M. (**2020b**). "Analyzing the directivity patterns of human speakers," in *Proceedings of the 46th DAGA Congress*, pp. 1141–1144.
Pörschmann, C., and Arend, J. M. (**2021**). "Supplementary material for 'Investigating phoneme-dependencies of spherical voice directivity patterns'" [Data set], J. Acoust. Soc. Am. (Last viewed June 18, 2021).
Pörschmann, C., Arend, J. M., and Brinkmann, F. (**2019**). "Directional equalization of sparse head-related transfer function sets for spatial upsampling," IEEE/ACM Trans. Audio Speech Language Process. **27**(6), 1060–1071.
Postma, B. N. J., Demontis, H., and Katz, B. F. G. (**2017**). "Subjective evaluation of dynamic voice directivity for auralizations," Acta Acust. united Acust. **103**(2), 181–184.
Postma, B. N. J., Jouan, S., and Katz, B. F. G. (**2018**). "Pre-Sabine room acoustic design guidelines based on human voice directivity," J. Acoust. Soc. Am. **143**(4), 2428–2437.
Postma, B. N. J., and Katz, B. F. (**2016**). "Dynamic voice directivity in room acoustic auralizations," in *Proceedings of the 42nd DAGA Congress*, March 14–17, Aachen, Germany, pp. 352–355.
Rafaely, B. (**2015**). *Fundamentals of Spherical Array Processing* (Springer-Verlag, Berlin), p. 193.
Richter, J. G., Pollow, M., Wefers, F., and Fels, J. (**2014**). "Spherical harmonics based HRTF datasets: Implementation and evaluation for real-time auralization," Acta Acust. united Acust. **100**(4), 667–675.
Saunders, G. (**1790**). *Treatise on Theaters* (I. and J. Taylor, London), pp. 1–94.
Trendelenburg, F. (**1929**). "Contribution to the question of voice directivity," Z. Techn. Phys. **11**, 558–563.
Williams, E. G. (**1999**). *Fourier Acoustics—Sound Radiation and Nearfield Acoustical Holography* (Academic Press, London).
Wyatt, B. (**1813**). *Observation on the Design for the Theatre Royal, Drury Lane* (J. Taylor, London), pp. 1–70.
Xie, B. (**2009**). "On the low frequency characteristics of head-related transfer function," Chin. J. Acoust. **28**, 1–13.

6.2 Impact of Face Masks on Voice Radiation

Pörschmann, C., Lübeck, T., & Arend, J. M. (2020). *J. Acoust. Soc. Am.*, *148*(6), 3663–3670. https://doi.org/10.1121/10.0002853

JASA INVITED-PAPER

Impact of face masks on voice radiation[a)]

Christoph Pörschmann,[b)] Tim Lübeck,[c),d)] and Johannes M. Arend[c),e)]
Institute of Communications Engineering, TH Köln—University of Applied Sciences, Betzdorfer Straße 2, 50679 Cologne, Germany

ABSTRACT:
With the COVID-19 pandemic, the wearing of face masks covering mouth and nose has become ubiquitous all around the world. This study investigates the impact of typical face masks on voice radiation. To analyze the transmission loss caused by masks and the influence of masks on directivity, this study measured the full-spherical voice directivity of a dummy head with a mouth simulator covered with six masks of different types, i.e., medical masks, filtering facepiece respirator masks, and cloth face coverings. The results show a significant frequency-dependent transmission loss, which varies depending on the mask, especially above 2 kHz. Furthermore, the two facepiece respirator masks also significantly affect speech directivity, as determined by the directivity index (DI). Compared to the measurements without a mask, the DI deviates by up to 7 dB at frequencies above 3 kHz. For all other masks, the deviations are below 2 dB in all third-octave frequency bands. © *2020 Acoustical Society of America*.
https://doi.org/10.1121/10.0002853
(Received 14 August 2020; revised 30 October 2020; accepted 11 November 2020; published online 14 December 2020)
[Editor: James F. Lynch] Pages: 3663–3670

I. INTRODUCTION

In times of crisis, such as the COVID-19 pandemic, security measures affect human face-to-face communication in many different ways. Speech intelligibility is impaired by greater physical distances between interlocutors, resulting in increased masking by background noise and a lower direct-to-reverberant ratio (DRR) at the listener position. In addition, face masks covering the mouth and nose cause transmission loss, reducing the energy radiated by the speaker. Depending on the frequency characteristics of the transmission loss, this can affect speech intelligibility (Palmiero *et al.*, 2016; Radonovich *et al.*, 2010). As the transmission loss of the masks may also vary with respect to the direction of radiation, the DRR could be further affected if, for example, the masks attenuate frontal sound radiation more than lateral radiation. Furthermore, other effects (e.g., lip reading) also strongly affect speech intelligibility (Matthews *et al.*, 2002; Sumby and Pollack, 1954; Summerfield, 1992).

So far, the acoustic effects of face masks have only been investigated in a small number of studies. Radonovich *et al.* (2010) and Palmiero *et al.* (2016) evaluated the influence of respirator masks worn by healthcare workers on speech intelligibility. In these studies, the influence of masks on the speech transmission index (STI) was determined for certain room settings. Recently, Goldin *et al.* (2020) analyzed three medical masks and found a low-pass characteristic attenuating frequencies above 2 kHz.

To further investigate to what extent the acoustic effects of face masks affect voice radiation and consequently speech intelligibility, we present full-spherical voice directivity measurements of a dummy head with a mouth simulator covered with six masks of different types, i.e., medical masks, filtering facepiece respirator masks, and cloth face coverings. We analyze the transmission loss caused by the masks as well as their influence on directivity. Although it is already clear that the transmission loss caused by the masks can reduce speech intelligibility, it has not yet been investigated how face masks affect speech directivity and, therefore, possibly also speech intelligibility. Resonances of vibrating structures of the mask or cases where the mask is not completely closed at the sides may lead to reduced frontal sound radiation compared to the lateral or rear radiation, impairing speech directivity. Accordingly, when a speaker inside a room faces a listener, the reverberant energy caused by sound radiation in all directions is decreased less than the direct sound energy, resulting in a reduced DRR at the listener position and possibly reduced speech intelligibility. Furthermore, the influence on the directivity could become relevant if the interlocutors do not look at each other directly but turn their heads slightly and thus radiate speech laterally, as is often observed in conversations.

II. MEASUREMENT PROCEDURE

The directivity measurements were done with a HEAD acoustics HMS II.3 dummy head and mouth simulator (head width 14.0 cm; head height 22.5 cm; head length 20.0 cm). However, as a mouth simulator cannot reflect phoneme-dependent effects, there are distinct differences between the

[a)]This paper is part of a special issue on "COVID-19 PANDEMIC ACOUSTIC EFFECTS."
[b)]Electronic mail: christoph.poerschmann@th-koeln.de, ORCID: 0000-0003-0794-0444.
[c)]Also at: Technical University of Berlin, Audio Communication Group, Berlin, Germany.
[d)]ORCID: 0000-0003-2870-095X.
[e)]ORCID: 0000-0002-5403-4076.

radiation of a dummy head and the human voice. For example, the mouth opening's variable size plays an important role (Brandner *et al.*, 2020), as well as sound radiation of the nasal passage and the position of the voice excitation in the vocal tract. These phoneme-dependent effects are one important aspect of the dynamic voice directivity which has been analyzed in several studies, e.g., in Kocon and Monson (2018), Katz and D'Alessandro (2007), Monson *et al.* (2012), and Pörschmann and Arend (2020). While simulations could probably examine some aspects of voice radiation such as the mouth opening's variable size, other aspects such as the physical contact between the mask and the vibrating lips are much harder to analyze. Despite these differences, a dummy head's radiation generally covers typical spatial characteristics of human voice radiation (Halkosaari *et al.*, 2005) and therefore provides a good approximation. Accordingly, we assume that the influence of the masks on voice radiation is comparable for a dummy head and human voice radiation.

The measurements were performed in the anechoic chamber of TH Köln, sized $4.5\,\text{m} \times 11.7\,\text{m} \times 2.30\,\text{m}$ (W × D×H), with a lower boundary frequency of about 200 Hz. As shown in Fig. 1, the HEAD acoustics HMS II.3 dummy head and mouth simulator was mounted on the VariSphear measurement system (Bernschütz *et al.*, 2010), which rotated the dummy head along a virtual sphere to the respective direction. In each direction, the excitation signal was played back over the mouth simulator and captured at a fixed position at a distance of 2 m from the center of the dummy head with a Microtech Gefell M296S omnidirectional microphone. The loudspeaker (mouth simulator) of the HEAD acoustics HMS II.3 was driven by an Apart MB-150 amplifier. An RME Babyface was used as AD/DA converter and microphone amplifier. The excitation signal was an emphasized sine sweep (2^{18} samples at a sampling rate of 48 kHz, length of 5.5 s). Impulse responses were measured for 2702 directions on a 44th order Lebedev sampling grid (Lebedev, 1977). In post-processing, adaptive low-frequency extension for frequencies below 200 Hz was applied to the impulse responses (Xie, 2009). Subsequently, the frequency and phase response of the loudspeaker (mouth simulator) was compensated by inverse finite impulse response (FIR) filtering with a frontal impulse response measured without a mask at azimuth $\phi = 0°$ and elevation $\theta = 0°$, corresponding to a free-field equalization of the measurements. Finally, the impulse responses were truncated and windowed to a length of 128 samples at a sampling rate of 48 kHz.

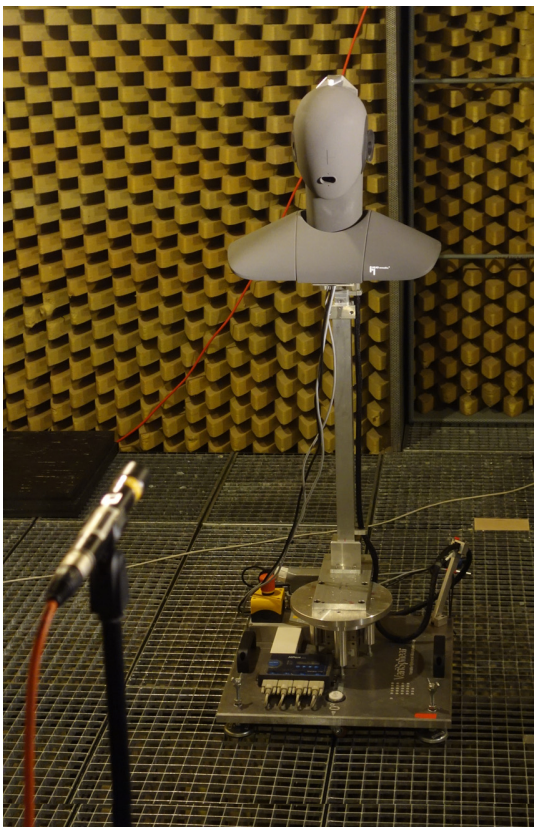


FIG. 1. (Color online) Measurement setup with the HEAD acoustics HMS II.3 dummy head and mouth simulator mounted on the VariSphear measurement system (Bernschütz *et al.*, 2010) and the omnidirectional Microtech Gefell M296S measurement microphone at a distance of 2 m from the dummy head.

Except for the reference directivity measurement of the mouth simulator without a mask, the measurement procedure was repeated twice for each of the following face masks:

(1) Disposable medical mask: Triple Layer Filter, dustproof, non woven earloop, (manufactured in China, brandname: Arvin, Protection Class DS3).
(2) Three-dimensional respirator mask, Protection Class KN 95 (manufacturer: Suzhou Jinruida Protective Equipment Co, Inc.).
(3) Fine dust respirator mask, MB 21, Protection Class FFP 2 (manufacturer: MB Filter Products AB, Sävedalen, Sweden).
(4) Microfibre scarf (manufacturer: Rose, material: 100% polyester).
(5) Cloth face covering: Single layer cotton (manufacturer: Modeatelier Scharn, Engelskirchen, Germany).
(6) Cloth face covering: Hand-made with two layers of cotton (manufacturer: Stoffliebe, Gelsenkirchen, Germany).

According to the World Health Organization (2020), these masks (or more generally mouth and nose covers) can be categorized as medical masks (1), filtering facepiece respirator masks (2, 3), or cloth face coverings (4, 5, 6). Figure 2 shows the dummy head with the six tested masks.

III. RESULTS

Every mask was measured twice. In between, it was taken off and put on again. We calculated the signed difference between both measurements in dB averaged over all directions for frequencies below 8 kHz to analyze any possible deviations. The deviations were only about 0.5 to 0.8 dB, so we did not further investigate the effects of putting on and taking off the mask and simply refer to the first

https://doi.org/10.1121/10.0002853

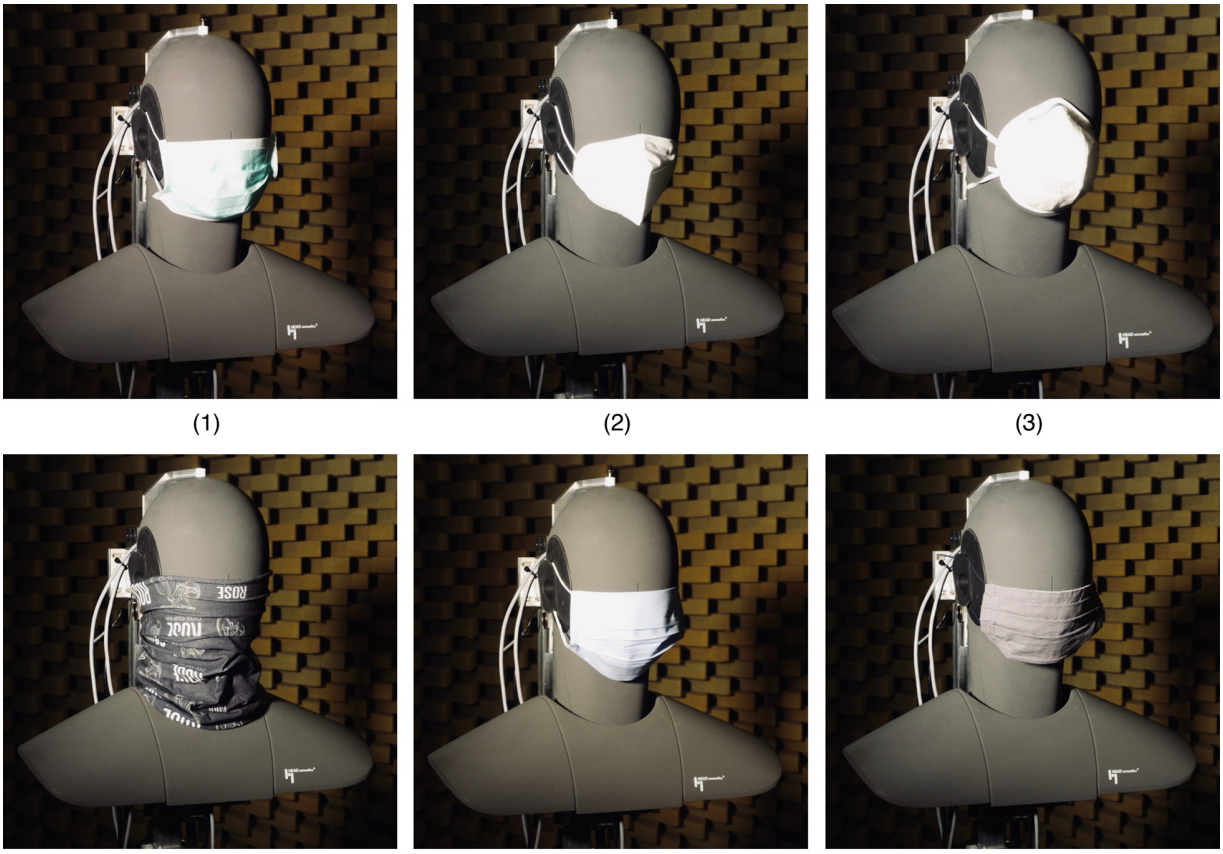


FIG. 2. (Color online) HEAD acoustics HMS II.3 dummy head and mouth simulator with all the masks tested.

measurements in the following. However, the publicly available dataset contains all measurements.[1]

A. Transmission loss

In the remainder, the directivity is denoted as $p(\Omega, \omega)$ with respect to the direction Ω (ϕ, θ) and the angular frequency $\omega = 2\pi f$, where f is the temporal frequency. ϕ denotes the azimuth angle ranging from -180° to $+180^\circ$, and θ the elevation ranging from -90° to 90°, where 90° is at the top, and -90° at the bottom. The transmission loss $\mathrm{T}(\Omega, \omega)$ between the reference directivity $p_{\mathrm{ref}}(\Omega, \omega)$ without a mask, and the directivity $p_{\mathrm{mask}}(\Omega, \omega)$ with a mask, can be expressed as

$$\mathrm{T}(\Omega, \omega) = 10\lg \frac{|p_{\mathrm{ref}}(\Omega, \omega)|^2}{|p_{\mathrm{mask}}(\Omega, \omega)|^2}. \tag{1}$$

Figure 3 (left) shows the results for frontal sound radiation $\Omega_0 = (\phi = 0^\circ, \theta = 0^\circ)$. Up to 2 kHz, the transmission loss is relatively low and flat for most masks but increases rapidly for masks 2 and 3 above 2 kHz with peaks of about 15 dB between 3 kHz and 5 kHz. Mask 6 shows a completely different frequency-dependent shape of the transmission loss. It exhibits a similar increase above 2 kHz, but already relevant transmission loss at low frequencies and a first strong peak at about 900 Hz.

For further analysis, we calculated $\mathrm{T}_{\mathrm{avg}}(\omega)$ by averaging the transmission loss over all radiation directions, Q,

$$\mathrm{T}_{\mathrm{avg}}(\omega) = 10\lg \frac{1}{Q} \sum_{q=1}^{Q} \frac{|p_{\mathrm{ref}}(\Omega_q, \omega)|^2}{|p_{\mathrm{mask}}(\Omega_q, \omega)|^2}. \tag{2}$$

Figure 3 (right) shows the results for the average transmission loss. In general, a similar trend can be observed as for the frontal sound radiation. However, for the medical respirator masks (masks 2 and 3), the peaks vary slightly with the radiation direction, and the shape of the curves becomes smoother and has less prominent peaks due to the averaging over direction. In contrast, for the other masks, the differences between Fig. 3, left and right, are much smaller, indicating that the peaks are rather independent of direction. Masks 1, 4, 5, and 6 show two peaks at about 900 Hz and 4 kHz, whereas for mask 4 the peaks are below 3 dB, and for masks 1 and 5, they are below 6 dB. Mask 6 exhibits peaks of more than 10 dB. Furthermore, for mask 6, the transmission loss is already above 3 dB at low frequencies. This may be due to the structure of this mask consisting of two layers of thick cloth.

B. Directivity analysis

To analyze the directivity in the horizontal and vertical plane, the 2702 measured impulse responses were

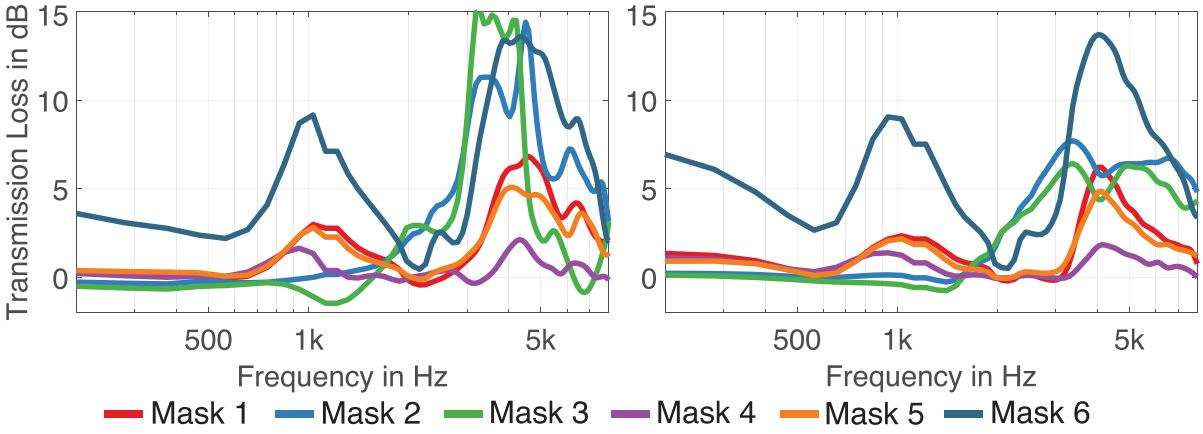


FIG. 3. (Color online) Transmission loss $T(\Omega_0, \omega)$, 1/12 octave smoothed, of the different masks for frontal sound radiation $\Omega_0 = (\phi = 0°, \theta = 0°)$ (left), and averaged over all directions $T_{avg}(\omega)$ (right).

transformed to the spherical harmonics (SH) domain (Williams, 1999) at an SH order $N = 35$ and resampled to 360 directions in 1° steps along the horizontal plane [$\phi = -180°$ to 180° (where positive angles pointing to the left), $\theta = 0°$], and along the vertical plane [$\phi = 0°, \theta = -180°$ to 180° (where 0° points to the front, 90° to the top, and 180° to the back)] using the inverse SH transform. Figures 4 and 5 show polar plots of the horizontal and vertical directivities in third-octave bands. We only present the directivity patterns for the frequency bands above 500 Hz, as we could not observe any relevant influence of the masks on directivity for lower frequencies. For the third-octave bands up to 1.25 kHz, the directivities vary between the masks less than 1 dB, and thus, barley differ from the reference (black curve). Only for the directivity

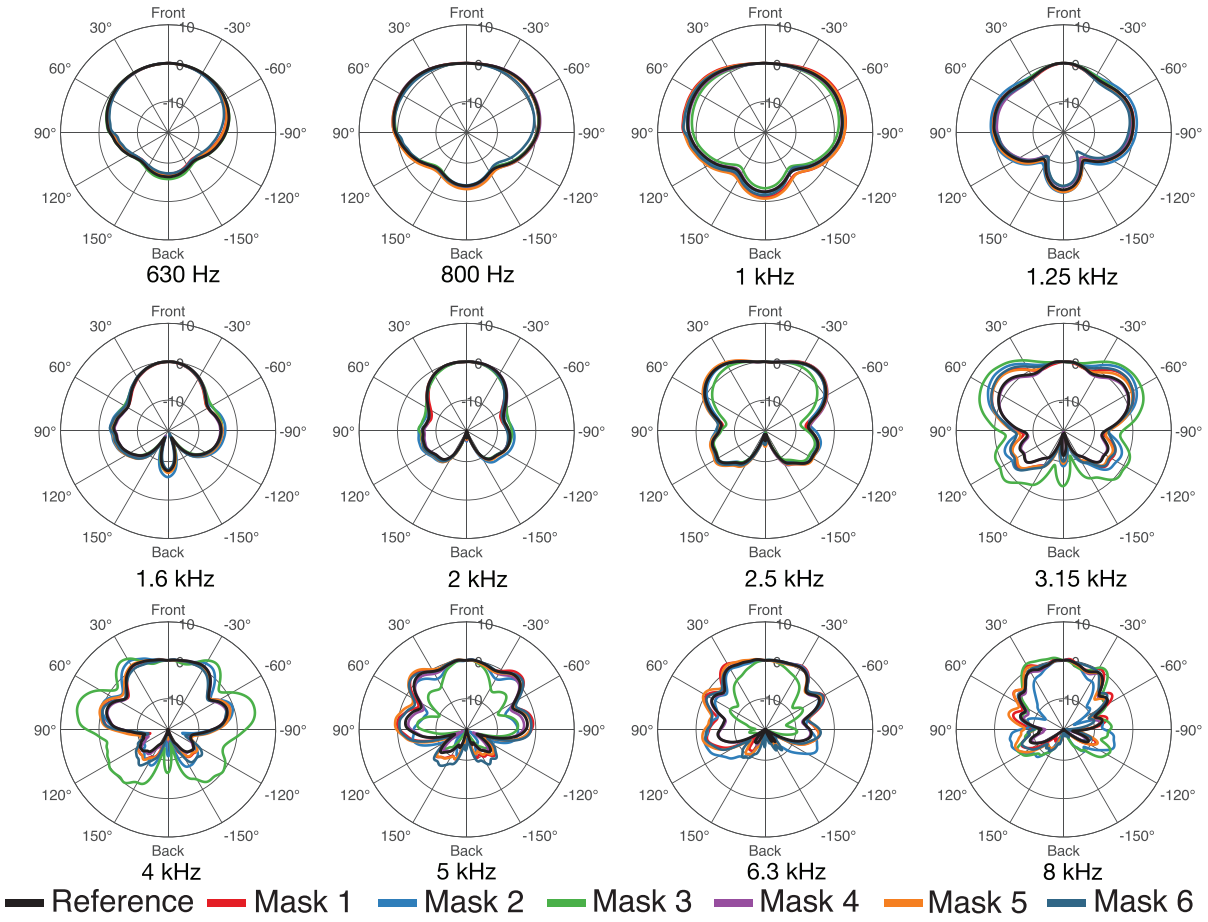


FIG. 4. (Color online) Directivity patterns in the horizontal plane ($\phi = -180°$ to 180°, $\theta = 0°$) for all masks in third-octave bands.

https://doi.org/10.1121/10.0002853

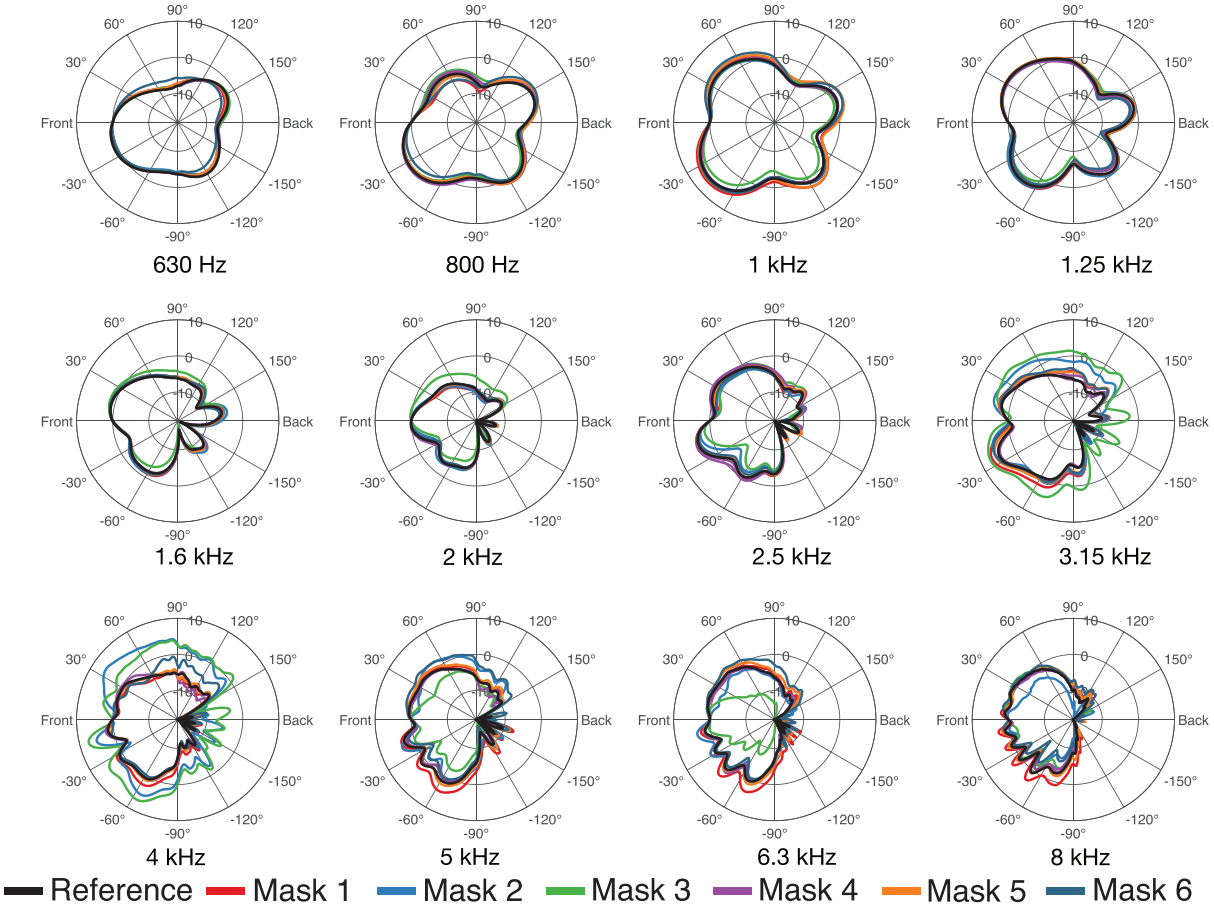


FIG. 5. (Color online) Directivity patterns in the vertical plane ($\phi = 0°, \theta = -180°$ to $180°$) for all masks in third-octave bands.

of mask 6 can some differences be observed, e.g., of 1 dB in the 630 Hz band at 60° in the vertical plane (Fig. 5). Above 1.25 kHz, the differences rise in magnitude with increasing frequency. As with the transmission loss, the filtering facepiece respirator masks (masks 2 and 3) exhibit the most prominent differences to the reference. In the 3.15 kHz band, which covers the same frequency range for which we observed the maximum transmission loss for the two masks, significant directivity variations occur. Furthermore, the directivities of masks 2 and 3 vary strongly for adjacent frequency bands. For example, the directivity at 4 kHz is much broader for mask 3 than for the other masks, while it is directed stronger to the front at 5 kHz. Please refer to the Appendix for a further illustration of the influence of the masks on sound radiation in the horizontal and vertical plane with respect to frequency.

C. Directivity index (DI)

For a more detailed analysis of the directivity, we determined the DI for spherical sound radiation, which can be calculated as

$$DI = 10\lg \frac{4\pi |p(\Omega_m, \omega)|^2}{\int_{-\pi}^{\pi}\int_{-\pi/2}^{\pi/2} |p(\Omega, \omega)|^2 \cos\theta \mathrm{d}\theta \mathrm{d}\phi}, \quad (3)$$

TABLE I. DI in the third-octave bands related to the frontal direction Ω_m.

Mask type	630 Hz	800 Hz	1 kHz	1.25 kHz	1.6 kHz	2 kHz	2.5 kHz	3.15 kHz	4 kHz	5 kHz	6.30 kHz	8 kHz
No mask	3.22	1.11	–0.42	1.86	4.99	6.60	3.72	3.54	4.97	4.00	4.47	5.69
Mask 1	3.43	1.18	–1.35	1.54	4.76	6.78	3.96	2.53	4.57	2.38	3.20	4.82
Mask 2	3.52	1.40	–0.44	1.18	4.55	6.50	4.21	0.42	0.93	4.12	4.93	7.21
Mask 3	3.23	1.26	0.48	2.09	4.75	6.14	4.88	–2.55	–2.29	7.86	9.65	6.47
Mask 4	3.38	0.93	–0.36	2.20	5.04	6.60	3.47	3.85	5.05	4.41	4.63	6.11
Mask 5	3.36	1.07	–1.22	1.58	4.69	6.57	3.65	2.79	4.43	2.80	3.33	5.22
Mask 6	3.55	1.46	–1.10	1.84	4.66	6.52	3.99	3.63	4.95	2.53	3.39	5.79

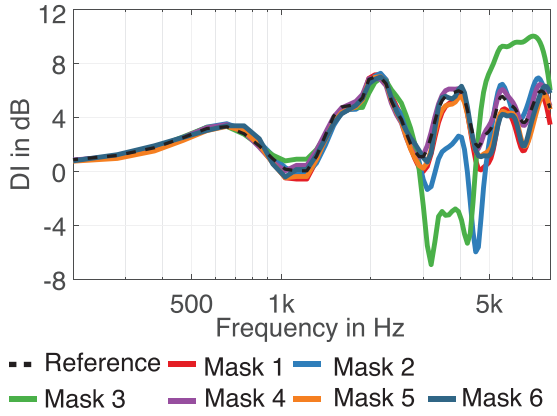


FIG. 6. (Color online) DI, 1/12 octave smoothed, of the different masks.

where Ω_m denotes the main direction of sound radiation. Usually, the direction with maximum magnitude of radiation of the human mouth is inclined slightly downwards (Marshall and Meyer, 1985), which can also be seen in Fig. 5. However, we compensated the measurements for the frontal direction (Ω_0), and furthermore, the frontal direction is included in all directivity plots. Therefore, we decided to calculate the DIs related to the main direction $\Omega_m = \Omega_0$. Table I summarizes the DIs for the third-octave bands between 630 Hz and 8 kHz. Apart from masks 2 and 3, the differences of the DI to the reference remain below 1 dB for frequencies up to 4 kHz and below 2 dB for frequencies up to 8 kHz. Again, the two facepiece respirator masks, masks 2 and 3, have the largest influence. In particular, between 3.15 and 6.3 kHz, the DI varies significantly. The differences are most prominent in the 4 kHz band, exceeding 4 dB (mask 2) and 7 dB (mask 3). This is supported by Fig. 6, which shows the DI with respect to frequency. In the frequency range from 3 to 5 kHz, the DI is significantly decreased by masks 2 and 3. At about 5 kHz the DI of mask 3 sharply rises towards high frequencies and is about 5 dB higher than the DI of all other measured datasets. The deviations of the DIs for the two facepiece respirator masks (masks 2 and 3) match the observations from Sec. III B, where we found the largest directivity variations for the same masks at the same frequencies (Figs. 4 and 5), and also the findings from Sec. III A, where we found the most considerable transmission loss for the corresponding masks in similar frequency regions.

IV. CONCLUSION

This paper presented measurements and analyses of sound radiation from a dummy head and mouth simulator wearing various face masks, i.e., medical masks, filtering facepiece respirator masks, and cloth face coverings. As discussed in Sec. II, a dummy head reproduces the typical spatial structure of human voice radiation, even though several aspects of human voice radiation like, e.g., the dynamic directivity, cannot be considered.

All examined masks result in a noticeable transmission loss at frequencies above 2 kHz. Thus, the transmission loss of the masks affects relevant frequency components of speech transmission, and therefore certainly impairs speech intelligibility. While in the frontal direction for the facepiece respirator masks (masks 2 and 3) the transmission loss increases strongly above 3 kHz by up to 15 dB, for the medical mask and most of the cloth face coverings, the transmission loss remains below 6 dB for frequencies above 3 kHz.

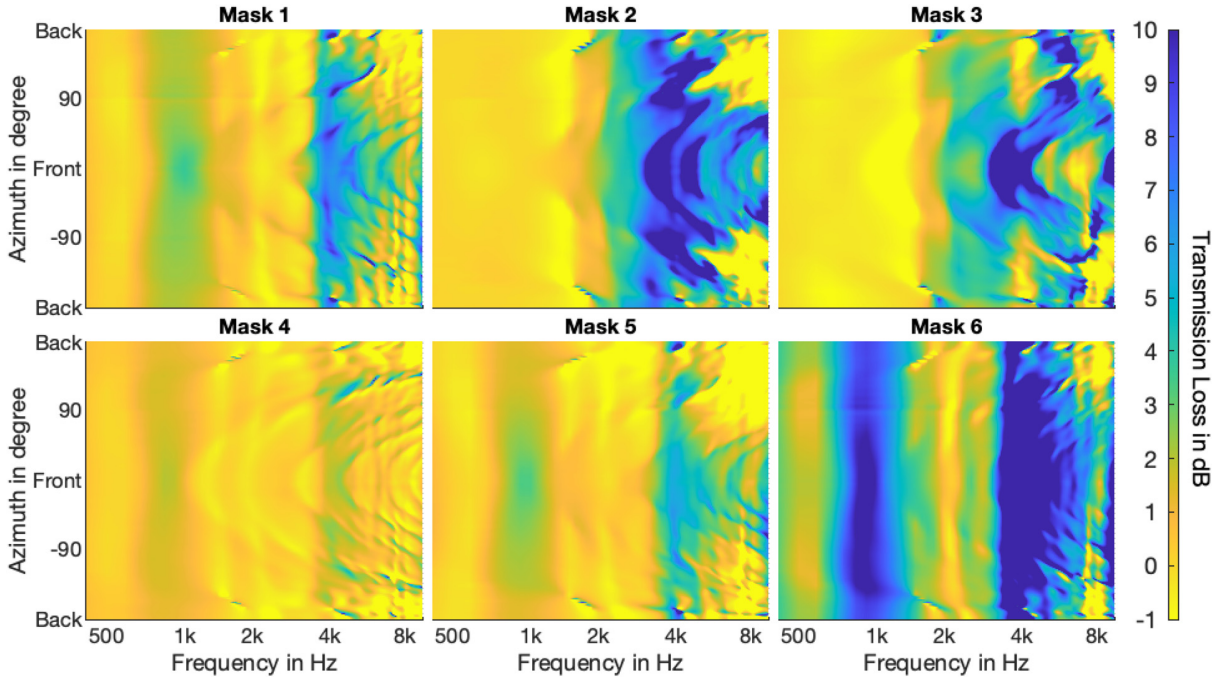


FIG. 7. (Color online) Transmission loss $T(\omega)$ in the horizontal plane with respect to frequencies from 400 Hz to 10 kHz.

https://doi.org/10.1121/10.0002853

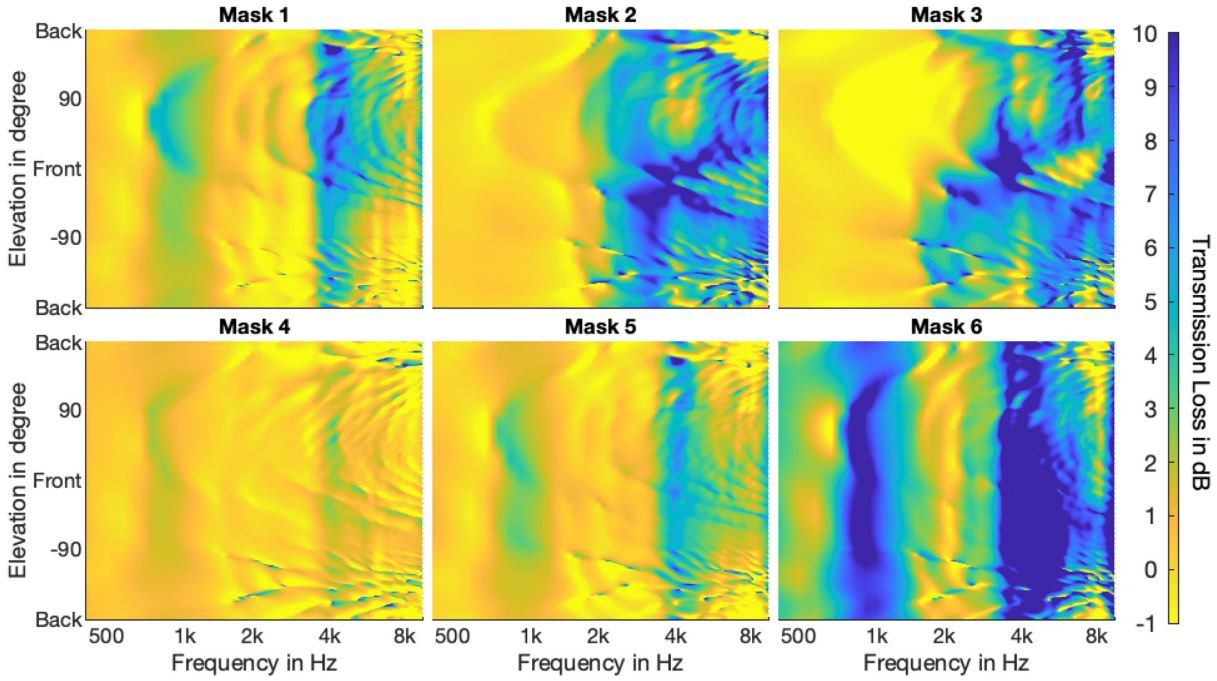


FIG. 8. (Color online) Transmission loss $T(\omega)$ in the vertical plane with respect to frequencies from 400 Hz to 10 kHz.

An exception is a hand-made cloth face covering (mask 6), which consists of two thick layers of cotton, showing a much higher transmission loss already at low frequencies. Our results are in line with the results of Goldin *et al.* (2020). The study analyzed three different masks and determined an attenuation from 4 to 6 dB for a simple medical mask and an attenuation of 12 dB up to 18 dB for the higher protective N95 masks in the frequency range between 3 and 7 kHz.

The analysis of the directivity showed that the masks affect speech directivity differently. Both tested facepiece respirator masks lead to an increase of the DI by up to 7 dB between 3 and 5 kHz, and one of them to an increased DI for 5 dB above 5 kHz. The other masks show only a weak influence on the directivity, affecting the DI by a maximum of 1 dB for frequencies up to 4 kHz and 2 dB for frequencies up to 8 kHz. The DI can be directly related to the DRR, for which Larsen *et al.* (2008) determined just-noticeable differences (JNDs) of about 2–3 dB in rooms with a DRR of 0 or +10 dB and JNDs of about 6–8 dB in rooms with a DRR of −10 or +20 dB. For some frequency bands, the change of the DI caused by the facepiece respirator masks is in the range of the JND. Therefore, the facepiece respirator masks might impair speech intelligibility in rooms. However, as the DRR is a broadband measure, it is hard to predict how a decrease of the DI in specific frequency bands affects speech intelligibility. It can be assumed that especially phonemes incorporating significant energy in the affected frequency range might show degraded intelligibility. The datasets from this study can be applied in follow-up studies similar to Fogerty *et al.* (2020) or Kokabi *et al.* (2019) using room acoustic simulations or virtual acoustics to analyze in more detail which way the masks affect speech intelligibility.

ACKNOWLEDGMENTS

C.P. and T.L. contributed equally to this work.

APPENDIX

Figures 7 and 8 show the transmission loss with respect to frequency in the horizontal and vertical plane for a frequency range of 400 Hz to 10 kHz. The figures illustrate the directional dependency of the transmission loss and support the observations from Secs. III A and III C. In particular, masks 2, 3, and 6 exhibit notable transmission loss, as indicated by the deep blue color.

[1]All directivity datasets are available in SOFA format under a Creative Commons CC BY-SA 4.0 license and can be downloaded at https://doi.org/10.5281/zenodo.3952320. The datasets will be published with a DOI on https://zenodo.org after acceptance of the paper.

Bernschütz, B., Pörschmann, C., Spors, S., and Weinzierl, S. (**2010**). "Entwurf und Aufbau eines variablen sphärischen Mikrofonarrays für Forschungsanwendungen in Raumakustik und Virtual Audio" ("Design and setup of a variable spherical microphone array for research applications in room acoustics and virtual audio"), in *Proceedings of the 36th DAGA*, March 15–18, Berlin, Germany, pp. 717–718.

Brandner, M., Blandin, R., Frank, M., and Sontacchi, A. (**2020**). "A pilot study on the influence of mouth configuration and torso on singing voice directivity," J. Acoust. Soc. Am. **148**(3), 1169–1180.

Fogerty, D., Alghamdi, A., and Chan, W.-Y. (**2020**). "The effect of simulated room acoustic parameters on the intelligibility and perceived reverberation of monosyllabic words and sentences," J. Acoust. Soc. Am. **147**(5), EL396–EL402.

Goldin, A., Weinstein, B., and Shiman, N. (**2020**). "How do medical masks degrade speech perception," Hear. Rev. **27**(5), 8–9.
Halkosaari, T., Vaalgamaa, M., and Karjalainen, M. (**2005**). "Directivity of artificial and human speech," J. Audio Eng. Soc. **53**(7–8), 620–631.
Katz, B., and D'Alessandro, C. (**2007**). "Directivity measurements of the singing voice," in *Proceedings of the 19th International Congress on Acoustics*, September 2–7, Madrid, Spain, pp. 2–7.
Kocon, P., and Monson, B. B. (**2018**). "Horizontal directivity patterns differ between vowels extracted from running speech," J. Acoust. Soc. Am. **144**(1), EL7–EL12.
Kokabi, O., Brinkmann, F., and Weinzierl, S. (**2019**). "Prediction of speech intelligibility using pseudo-binaural room impulse responses," J. Acoust. Soc. Am. **145**(4), EL329–EL333.
Larsen, E., Iyer, N., Lansing, C. R., and Feng, A. S. (**2008**). "On the minimum audible difference in direct-to-reverberant energy ratio," J. Acoust. Soc. Am. **124**(1), 450–461.
Lebedev, V. I. (**1977**). "Spherical quadrature formulas exact to orders 25–29," Sib. Math. J. **18**(1), 99–107.
Marshall, A. H., and Meyer, J. (**1985**). "The directivity and auditory impressions of singers," Acustica **58**, 130–140.
Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., and Harvey, R. (**2002**). "Extraction of visual features for lipreading," IEEE Trans. Pattern Anal. Mach. Intell. **24**(2), 198–213.
Monson, B. B., Hunter, E. J., and Story, B. H. (**2012**). "Horizontal directivity of low- and high-frequency energy in speech and singing," J. Acoust. Soc. Am. **132**(1), 433–441.
Palmiero, A. J., Symons, D., Morgan, J. W., and Shaffer, R. E. (**2016**). "Speech intelligibility assessment of protective facemasks and air-purifying respirators," J. Occup. Environ. Hyg. **13**(12), 960–968.
Pörschmann, C., and Arend, J. M. (**2020**). "Analyzing the directivity patterns of human speakers," in *Proceedings of the 46th DAGA*, March 16–19, Hannover, Germany, pp. 1141–1144.
Radonovich, L. J., Yanke, R., Cheng, J., and Bender, B. (**2010**). "Diminished speech intelligibility associated with certain types of respirators worn by healthcare workers," J. Occup. Environ. Hyg. **7**(1), 63–70.
Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**(2), 212–215.
Summerfield, Q. (**1992**). "Lipreading and audio-visual speech perception," Philos. Trans. R. Soc. London B Biol. Sci. **335**(1273), 71–78.
Williams, E. G. (**1999**). *Fourier Acoustics—Sound Radiation and Nearfield Acoustical Holography* (Academic Press, London, UK).
World Health Organization (**2020**). "Advice on the use of masks in the context of COVID-19: Interim guidance, 5 June 2020," https://apps.who.int/iris/handle/10665/332293 (Last viewed December 1, 2020).
Xie, B. (**2009**). "On the low frequency characteristics of head-related transfer function," Chin. J. Acoust. **28**(2), 116–128.

6.3 A Method for Spatial Upsampling of Voice Directivity by Directional Equalization

Pörschmann, C., & Arend, J. M. (2020). *J. Audio Eng. Soc.*, *68*(9), 649–663. https://doi.org/10.17743/jaes.2020.0033

C. Pörschmann and J. M. Arend, "A Method for Spatial Upsampling of Voice Directivity by Directional Equalization"
J. Audio Eng. Soc., vol. 68, no. 9, pp. 649–663, (2020 September.).
DOI: https://doi.org/10.17743/jaes.2020.0033

A Method for Spatial Upsampling of Voice Directivity by Directional Equalization

CHRISTOPH PÖRSCHMANN,[1] *AES Associate Member,* **AND**
(Christoph.Poerschmann@th-koeln.de)
JOHANNES M. AREND,[1,2] *AES Student Member*
(Johannes.Arend@th-koeln.de)

[1]*Institute of Communications Engineering, TH Köln - University of Applied Sciences, 50679 Cologne, Germany*
[2]*Audio Communication Group, Technical University of Berlin, 10587 Berlin, Germany*

To describe the sound radiation of the human voice into all directions, measurements need to be performed on a spherical grid. However, the resolution of such captured directivity patterns is limited and methods for spatial upsampling are required, for example by interpolation in the spherical harmonics (SH) domain. As the number of measurement directions limits the resolvable SH order, the directivity pattern suffers from spatial aliasing and order-truncation errors. We present an approach for spatial upsampling of voice directivity by spatial equalization. It is based on preprocessing, which equalizes the sparse directivity pattern by spectral division with corresponding directional rigid sphere transfer functions, resulting in a time-aligned and spectrally matched directivity pattern that has a significantly reduced spatial complexity. The directivity pattern is then transformed into the SH domain, interpolated to a dense grid by an inverse spherical Fourier transform and subsequently de-equalized by spectral multiplication with corresponding rigid sphere transfer functions. Based on measurements of a dummy head with an integrated mouth simulator, we compare this approach to reference measurements on a dense grid. The results show that the method significantly decreases errors of spatial undersampling and this allows a meaningful high-resolution voice directivity to be determined from sparse measurements.

0 INTRODUCTION

In many everyday situations, we experience the influence of human voice directivity. Loudness and timbre of a human speaker change significantly when facing us or turning away from us. Often we use the directivity intuitively, for example when facing a person in meetings or casual conversations. Already in 1790, Saunders [1] observed in his studies that depending on the direction, speech could be perceived by a listener up to different distances to the speaker. This study was confirmed by Wyatt [2] and Henry [3] and can be regarded as a first indirect measurement of voice directivity in the horizontal plane. First direct measurements of human voice directivity were carried out more than 90 years ago by Trendelenburg [4], determining patterns for several vocals and fricatives in the horizontal plane. Later, Dunn and Farnsworth [5] measured directivity patterns at different distances for a spoken sentence in third-octave bands from 63 Hz up to 12 kHz. Other investigations were based on dummy heads with integrated mouth simulators [6] or compared dummy head measurements to simple models based on a point source located on a rigid sphere [7–9]. Halkosaari et al. [8] revealed differences of more than 10 dB between the voice directivity of a human speaker and a dummy head for broadband speech. The study indicated that the directivity patterns of the mouth simulators were too strong at frontal directions and high frequencies.

Human voice directivity data are also relevant to realize immersive teleconferencing systems or more generally to integrate actors or singers in virtual reality (VR) and augmented reality (AR) applications. When reproducing one's own voice in a virtual acoustic environment to investigate the perception of self-generated sound, human voice directivity also plays an essential role [10–12]. One specific aspect of the human voice directivity is its dynamic directivity, i.e., time-variant alterations that occur while articulating [13, 14] or singing [15]. Most auralizations of directivity in virtual acoustic environments assumed a time-invariant di-

rectivity [10, 16]. However, in more recent studies, Postma et al. [17, 18] analyzed the relevance of dynamic voice directivity for virtual acoustic environments. The results of these studies indicate that auralizations involving dynamic voice directivity are perceived as more plausible and exhibit a wider apparent source width than auralizations with static voice directivity or omnidirectional radiation.

To adequately determine human voice directivity while speaking or singing, the sound radiation needs to be captured for an appropriately large number of directions. Early measurements of directivity patterns were carried out separately for each frequency band [5] and often the data were acquired in specific planes only, either as in [19] in the horizontal plane or as in [6, 20, 21] in the horizontal and vertical plane. However, for reproduction in VR or AR systems, measurements on spherical sampling grids are advantageous. So far only a limited number of investigations on spherical datasets have been published, e.g., [22, 23, 13]. In the study of Kob [22], the focus was laid on the development of an artificial singer, and thus no detailed values are given. Brandner et al. [13] published only data and plots for the horizontal and vertical plane. Detailed values can be obtained from the study of Chu and Warnock [23] providing average values and standard deviations of measurements for 40 subjects and a B&K Head and Torso Simulator in third-octave bands for 92 positions measured on a spherical grid.

In general, voice directivity measurements can be performed either sequentially for an arbitrary number of directions or simultaneously using a surrounding microphone array. In the case of sequential measurements, time-variant aspects, which are for example caused by articulation-dependent mouth opening shapes, cannot be resolved and are as in [23] often not considered. As it is nearly impossible for subjects to perform exactly the same articulations for numerous measurements, capturing time-variant aspects of the human directivity with sequential measurements is hardly feasible. As a consequence, only the time-averaged spectrum in specific frequency bands is analyzed in the related studies. Other studies relying on sequential measurements were applied using dummy heads, which generally cannot reflect time-variant features, e.g., [9]. When applying a surrounding microphone array [24, 25, 13, 11] for simultaneous measurements, the complete capturing is done at once, and thus, loudness or articulation-dependent properties of human voice directivity can be investigated. However, as the setup of such surrounding arrays is restricted to a limited number of microphones, the measured dataset shows a low spatial resolution. Accordingly, methods are required for spatial upsampling of sparsely measured directivity patterns by an appropriate interpolation between the measurement directions. To the best of our knowledge, however, no scientific studies on spatial upsampling of voice directivity measurements have been published so far, so the present study can be regarded as the first approach to this matter.

Applying the principle of reciprocity [26], both head-related transfer functions (HRTFs) and voice directivity patterns can be formulated as an acoustic radiation problem [27, 28]. Accordingly, spatial upsampling of voice directivity obtained from measurements with a low spatial resolution can be performed comparably to spatial upsampling of HRTF sets, for which many methods have been elaborated, (e.g., [29–31]). Spatial upsampling of HRTF sets is nowadays often performed using a transformation into the spherical harmonics (SH) domain [32, ch. 6], [33, ch. 1]. In this case, sets of HRTFs measured on a spherical sampling grid are decomposed into spherical base functions of different spatial order N by applying a spherical Fourier transform (SFT). As these SH base functions are spatially continuous, upsampling can be performed by an inverse SFT on a dense sampling grid [34]. However, if the measurement grid was sparse, the coefficients obtained from the SFT involve so-called sparsity errors due to order truncation and spatial aliasing [35, 36]. To reduce these errors, several authors performed a time alignment of the HRTFs by eliminating their linear phase components before the SFT [34, 37–39]. By this, the ear position is virtually shifted to the center of the head, which reduces the spatial complexity of the dataset and as a consequence decreases sparsity errors. In this context, we recently presented the Spatial Upsampling by Directional Equalization (SUpDEq) method [40] and proposed a spatial equalization of HRTFs with corresponding rigid sphere transfer functions (STFs). The STFs can be regarded as a dataset featuring basic temporal and spectral components but leaving out information on the fine structure of the head.

In [41], we presented a first approach to apply the method to a sequentially measured radiation pattern of a dummy head with a mouth simulator. In the present paper, we extend these investigations and analyze the suitability of the SUpDEq method for human voice directivity. As no datasets or methods to measure human voice directivites on a dense grid are available, we rely in our study on measurements with a dummy head and mouth simulator, which allows us to compare the interpolation of simultaneous measurements carried out with a surrounding microphone array to a high-resolution dataset obtained from sequential measurements. Thus, the approach forms a basis for subsequent studies on human voice directivity that include the analysis of time-variant and articulation-dependent effects.

The paper is organized as follows: Sec. 1 presents the method for the spatial upsampling of sparsely measured voice directivity, followed in Sec. 2 by a description of the measurement devices, procedures, and datasets that were used in this study. Sec. 3 analyzes the measurements depending on the grid type and measurement procedure. Sec. 4 summarizes the results of this evaluation and discusses which specific issues need to be considered when measuring directivity of the human voice instead of that of a dummy head. Finally, in Sec. 5, we draw a conclusion and give an outlook on applications that can benefit from the presented method.

1 METHOD

A directivity can be determined by the pressure $p(\omega, \Omega_g)$, measured at a defined distance for G directions $\Omega_g = \{(\phi_1, \theta_1), \ldots, (\phi_G, \theta_G)\}$ at azimuth ϕ and elevation θ as

a function of the angular frequency ω. Alternatively, the directivity can be described in the SH domain benefiting from the spatial continuity and the orthogonality of the SH basis functions. This approach, which is novel for measured voice directivity, is commonly used for HRTFs [33, 36, 42, 40] and has been applied to directivity measurements of musical instruments as well [43].

In the SH domain, directivity is presented by SH coefficients that are calculated by applying complex base functions $Y_n^m(\Omega)$ of order n and degree m with respect to the angular direction Ω [32, 33]. The inverse SFT is used to recover p at arbitrary directions. In case the spatial resolution of the directivity is limited to a maximum order of N, the (discrete) inverse SFT can be formulated in matrix form [33]

$$p = Y p_{nm}, \tag{1}$$

with $p = [p(\omega, \Omega_1), \ldots, p(\omega, \Omega_G)]$ the directivity measurements over the G directions and p_{nm} holding the directivity coefficients in the SH domain. Y represents the SFT transformation matrix, sized $G \times (N+1)^2$, defined by its g-th row

$$Y_g = [Y_0^0(\Omega_g), Y_1^{-1}(\Omega_g), \ldots, Y_N^N(\Omega_g)]. \tag{2}$$

The coefficients in the SH domain can then be determined by a multiplication of P and the pseudo-inverse of the transformation matrix Y. Accordingly, the (discrete) SFT yields

$$p_{nm} = Y^\dagger p. \tag{3}$$

The pseudo-inverse $Y\dagger$ can be determined by

$$Y^\dagger = (Y^H Y)^{-1} Y^H, \tag{4}$$

with $(\cdot)^H$ denoting the Hermitian operator. To avoid sparsity errors due to order truncation and spatial aliasing in the SFT, the directivity needs to be measured at a sufficiently high number of directions

$$G \geq (N+1)^2. \tag{5}$$

However, for directivity measurements, the number of directions is limited either by the number of sequential measurements or the setup of the surrounding microphone array. Consequently, the analyzed patterns are sparse and can only be determined up to a limited order in the SH domain, resulting in an imprecise description of the directivity. To entirely consider the spatial properties of a spherical head model, an order given by $N \geq kr$ is required, with $k = \omega/c$, r the head radius, and c the propagation speed of sound [44, 45]. Assuming $r = 8.75$ cm as the average human head radius [46] and $c = 343$ m/s, human voice radiation, which according to recent studies contains important information up to 16 kHz [47, 14], would require a minimum spatial order $N \approx 26$. However, this cannot be regarded as a fixed value, since differences of human voice radiation to a modeled point source located on a spherical head, e.g., caused by the specific head shape, the mouth opening size, or sound radiation of the nasal passage, might increase the required spatial order. Similarly, HRTFs show higher minimum required spatial orders than the spherical head model ($N \approx 40$ for frequencies up to 20 kHz [48]). However, the spherical head model provides a good starting point for estimating the spatial order required for both HRTF and directivity, since it already captures spatial phase changes quite well, which are largely responsible for the high spatial orders of the HRTF or directivity. Therefore, we assume that the minimum required spatial order for the directivity and HRTF is quite similar. If, however, the maximum spatial resolution of the array measurement is lower than that of p (i.e., the spatial resolution of the directivity in this case), order-truncation errors and spatial aliasing occur, resulting in an impaired amplitude and phase of the directivity set. For a detailed analysis of the various aspects of these sparsity errors please refer to [36, 42]. Several approaches have been carried out for HRTFs to reduce the spatial complexity of a spherical dataset by an appropriate alignment prior to the SFT [34, 38, 49], among them the SUpDEq method [40]. Based on directivity patterns of a dummy head we evaluate the SUpDEq method, which in subsequent studies shall be used to determine human voice directivity from measurements on sparse sampling grids.

Fig. 1 shows the basic structure of the SUpDEq method. In a first step, the sparsely measured directivity set p captured at S sampling directions $\Omega_s = \{(\phi_1, \theta_1), \ldots, (\phi_S, \theta_S)\}$ is equalized with an appropriate equalization dataset h_{EQ}, which holds an equalization function $h_{EQ}(\omega, \Omega_s)$ for each sampled direction:

$$p_{EQ}(\omega, \Omega_s) = \frac{p(\omega, \Omega_s)}{h_{EQ}(\omega, \Omega_s)} \tag{6}$$

The equalization dataset represents a simplified directivity of a spherical head model, which carries no information on the specific shape of the mouth opening or the form, e.g., of the cheekbones, but only features the basic shape of a spherical head. By the equalization the mouth is virtually shifted to the center of the head. This reduces energy in higher orders, which originates from rapid phase changes between neighboring directions for an off-center location of the mouth [50]. Accordingly, the directional dependencies of p are reduced and only the differences between the measured directivity dataset and the equalization dataset remain in p_{EQ}. As a result, artifacts are expected to be much smaller when p_{EQ} is transformed to the SH domain than transforming the non-equalized p. In this study rigid sphere transfer functions describing the sound radiation from a point source positioned on the surface of a rigid sphere into the far field are applied [33, p. 49] and the equalization dataset is calculated as:

$$h_{EQ,nm} = 4\pi \mathrm{i}^{-n} \Big[j_n(kr) - \frac{j_n'(kr)}{h_n^{(2)'}(kr)} h_n^{(2)}(kr) \Big] [Y_n^m(\Omega_e)]^*, \tag{7}$$

with r the head radius, Ω_e the mouth position, j_n the spherical Bessel function of the first kind, h_n the spherical Hankel function of the first kind, and $h_n^{(2)}$ the spherical Hankel function of the second kind, as well as the derivative of the respective functions marked by a $'$. As this equalization dataset is described analytically, it can be determined at a freely chosen maximum order, typically a high order $N \geq$

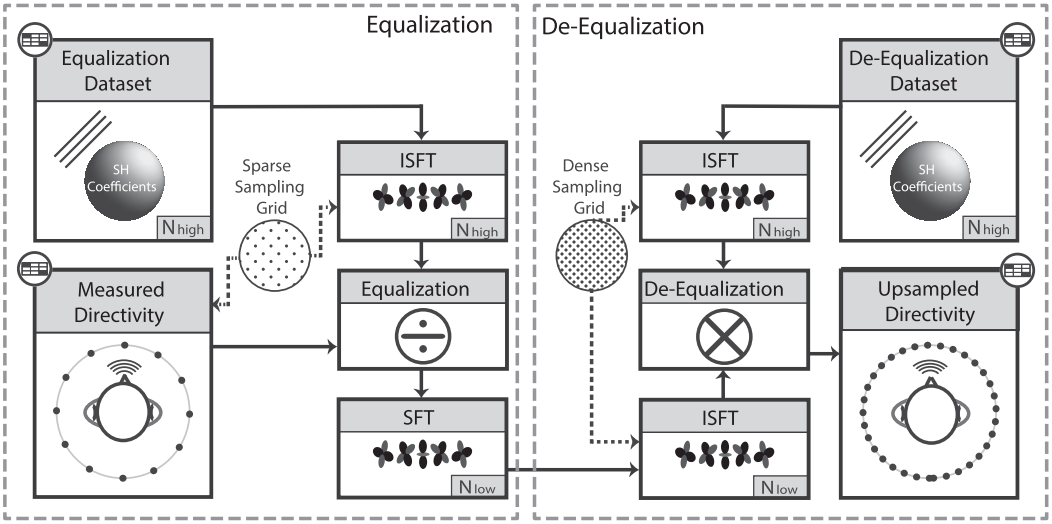


Fig. 1. Block diagram of the SUpDEq method for spatial upsampling of a voice directivity adapted from [40]. Left panel: a sparsely measured directivity is equalized on the corresponding sparse sampling grid. The set is then transformed to the SH domain with $N = N_{low}$. Right panel: the equalized set is de-equalized on a dense sampling grid, resulting in a dense spatially upsampled dataset.

32. The equalization dataset h_{EQ} is calculated by applying an inverse SFT according to Eq. (1) on the same sparse grid Ω_s for which the directivity was captured. According to Marshall [20], who measured a maximum sound radiation at an azimuth of $\phi = 0°$, slightly downward at an elevation between $\theta = -20°$ and $\theta = -30°$, we defined Ω_e as $\phi = 0°$ and $\theta = -25°$. The radius r corresponds to the radius of a human head and can be calculated according to Algazi et al. [51] in order to ensure an optimal spherical-head model.

Subsequently, the SH coefficients $p_{EQ,nm}$ are obtained by an SFT of the equalized directivity set p_{EQ}. The SFT according to Eq. (3) is performed at an appropriate low maximum order N_{low} satisfying Eq. (5). Then an upsampled directivity set p_{EQ} is calculated by applying an inverse SFT according to Eq. (1) on a dense sampling grid $\Omega_d = \{(\phi_1, \theta_1), \ldots, (\phi_D, \theta_D)\}$, with $D \gg S$. If N_{low} is chosen appropriately, sparsity errors are much smaller as for the unprocessed datasets. Finally, the directivity set is de-equalized by means of a spectral multiplication with a de-equalization dataset h_{DEQ}:

$$p_{DEQ}(\omega, \Omega_d) = p_{EQ}(\omega, \Omega_d) \cdot h_{DEQ}(\omega, \Omega_d)\,. \tag{8}$$

The de-equalization recovers energy at higher spatial orders that was removed in the first step. For de-equalization, again the rigid sphere transfer functions for sound radiation into the far field, sampled on a dense sampling grid Ω_d, are used.

If the directional equalization sufficiently removes all the energy in higher spatial orders above N_{low}, the upsampled directivity set completely reconstructs the radiation pattern of the sound source. Otherwise, deviations are caused by signal energy that, after the equalization, is still apparent at high spatial orders $N > N_{low}$. This results in spatial aliasing and order truncation (artifacts) as this signal energy is irreversibly mirrored to lower orders $N \leq N_{low}$ [45]. In Sec. 3 the influence of these deviations is analyzed. For this, we compare the SUpDEq method to datasets that are spatially upsampled in the SH domain, without any form of spatial equalization or de-equalization.

2 MATERIALS

All measurements of this study were performed in the anechoic chamber of TH Köln, sized 4.5 m × 11.7 m × 2.30 m (W × D × H), showing a lower boundary frequency of about 200 Hz. We captured the directivity with two different setups. The first was a sequential measurement system that enabled measurements on a dense grid. The second was a surrounding microphone array (SMA) shaped as a pentakis dodecahedron with 32 microphones, allowing for simultaneous measurements of the full spherical directivity in one step. Because human voice directivity patterns measured on a dense grid, which could serve as ground truth, are neither available nor could be easily obtained, in this study we examined the mouth directivity of a HEAD acoustics HMS II.3 dummy head. The dummy head has a head width of 14.0 cm, a head height of 22.5 cm, and a head length of 20.0 cm. From this data, according to Algazi et al. [51] an optimal head radius of $r = 8.8$ cm can be calculated, which is relevant when determining the rigid sphere transfer functions for equalization and de-equalization. The loudspeaker (mouth simulator) of the HEAD acoustics HMS II.3 was driven by an Apart MB-150 amplifier, and an RME Babyface audio interface was used as an AD / DA converter and microphone preamplifier. The excitation signal for all measurements was an emphasized sine sweep with +20 dB low-shelf at 100 Hz (2^{18} samples at a sampling rate of 48 kHz, length of 5.5 s). In general, the complete measurement procedure is comparable to the HRTF measurements described in [52]. To process the measured datasets, the SUpDEq toolbox [40] was used, utilizing routines from SOFiA toolbox [53] and AKtools [54] for SH signal processing.

(a) (b)

Fig. 2. Setups for measuring voice directivity of the HEAD acoustics HMS II.3 in the anechoic chamber of TH Köln. (a) Sequential measurements setup with the VariSphear device. (b) Simultaneous measurements setup with the surrounding microphone array (SMA).

2.1 Sequential Array Measurements

The sequential measurements were carried out applying the VariSphear measurement system [55] for precise positioning of the dummy head at the spatial sampling positions and for capturing the directivity. A Microtech Gefell M296S omnidirectional microphone was placed at a fixed position at a distance of 2 m from the center of the VariSphear. For the measurement, the dummy head was mounted on the robot arm of the VariSphear as shown in Fig. 2(a). In combination with the rotatable base plate, this setup allowed for full 3D rotation of the head on a virtual sphere. The whole measurement procedure was controlled by the Matlab-based VariSphear software. Besides the motor control and impulse response capture modules, the software provides automatic error detection to check every measured impulse response for noticeable variations with reference to previous measurements. This process ensured the validity of all obtained impulse responses.

We measured the directivity on a Lebedev [56] full spherical grid with 2,702 points and transformed it to the SH domain with $N = 35$, which allows for a nearly perfect interpolation of voice directivity. The SH representation of the dataset serves as a reference in all further analysis throughout this paper.

2.2 Surrounding Array Measurements

To compare the sequential to a simultaneous measurement procedure, we additionally captured the voice directivity of the HEAD acoustics HMS II.3 with a surrounding microphone array (SMA) [25, 11]. The approach using an SMA is generally not restricted to impulse response measurements and offers the possibility to study articulation-dependent and time-variant influences of the directivity. The basic shape of the array used in the present study is a pentakis dodecahedron with 32 Rode NT5 cardioid microphones located at the vertices on a constant radius of 1 m. This is at least sufficient for a spatial order of $N = 4$. For the presented measurements, an additional Rode NT5 microphone was placed at the frontal position as a reference for spectral equalization in postprocessing. Fig. 2(b) shows the array in the anechoic chamber of TH Köln with the dummy head inside. To minimize the differences to the sequential measurements, the dummy head was mounted on the same robot arm as for the VariSphear measurements. Four RME Octamic II devices served as a preamplifier and AD / DA converter for the 32 microphones of the SMA. All signals of the SMA were managed with two RME Fireface UFX audio interfaces (see [25, 11] for a more detailed description of the SMA setup). One of these audio interfaces was also used as a preamplifier and AD / DA converter for the reference microphone. Apart from this, we used the same setup as for the sequential measurements.

2.3 Postprocessing

In the subsequent postprocessing, which was based on the implementation and description in [52], all raw measurement data were first truncated and windowed. Then the frequency and phase response of the loudspeaker (mouth simulator) were compensated by inverse FIR filtering with the frontal impulse response ($\phi = 0^\circ$ and $\theta = 0^\circ$), which was either obtained as one sampling point of the dense grid (sequential measurements) or as the measurement with the reference microphone (simultaneous measurements). This results in a flat frequency response for frontal sound radiation. After postprocessing the final length of each impulse response is 128 samples at a sampling rate of 48 kHz.

2.4 Test Datasets

Based on the described measurements we created various sparse sets, serving as input data for the technical evaluation. Regarding the sequential measurements using the VariSphear, we generated datasets for two different sparse grid types from the measured reference set: Lebedev grids of different orders $N_{low} = 4, 7, 10, 13$ (corresponding to 38, 86, 170, and 266 sampling points) as well as a grid with the geometry of a pentakis dodecahedron with 32 sampling points (pentakis grid), which allows a direct comparison to the sparse sets measured with the SMA. The sparse sets were obtained from the SH representation of the reference dataset by applying the inverse SFT on the respective grids. Then, for each of the sparse grids, we created two different types of test datasets, which we used for the further analysis and which we upsampled to a Lebedev grid with 2,702 sampling points. The first type of test datasets was created applying the SUpDEq method to obtain one de-equalized set for each of the sparse datasets. The second type of datasets was obtained by an SH interpolation of the sparse sets without any further pre or postprocessing and is referred to in the following as unprocessed. Since unprocessed upsampling in the SH domain is the starting point for many approaches to improve spatial upsampling, it can be considered a basic state-of-the-art interpolation.

2.5 Distance Error Compensation

For the measurements with the SMA, a further processing step was required. Even though the setup and measurements were carried out with great care, small positioning errors remained, mainly caused by inaccuracies of the microphone positions of the SMA and off-center placement of the dummy head in the SMA. Both results were slightly varying distances from the center of the head to the microphones. Our analysis in [57] revealed that even small deviations lead to strong impairments in the spatial upsampling of HRTFs. In particular, we found that deviations in the range of $\pm 2\,cm$ almost completely void the benefits of SUpDEq. It can be assumed that these results are similar for voice directivity. Thus, to compensate these deviations, we applied a method that we already proposed and evaluated in [57].

The distance error compensation (DEC) benefits from the directional equalization of SUpDEq and can be briefly summarized as follows: the equalization removes direction-dependent spectral and temporal components from the measured directivity and results in (nearly) time-aligned impulse responses. Ideally after equalization only deviations due to positioning inaccuracies remain and thus the onset differences of the equalized impulse responses directly relate to the distance errors of the measurement setup. Thus we apply a simple onset-detection to the equalized impulse responses to estimate the distance errors and determine the required distance shift Δd. The distance shift itself is performed by a directional equalization (Eq. (6)) and subsequent de-equalization (Eq. (8)) at different distances in the same way as described in [58]. Instead of a plane wave as normally used for equalization and de-equalization (Eq. (7)), rigid sphere transfer functions representing the sound radiation of a point source located on a rigid sphere [33, p. 46] to a receiver at a distance d are utilized for the equalization and de-equalization dataset:

$$h_{DEC,nm}(d) = 4\pi(-\mathrm{i})kh_n^{(2)}(kd) \left[j_n(kr) - \frac{j_n'(kr)}{h_n^{(2)'}(kr)} h_n^{(2)}(kr)\right][Y_n^m(\Omega_e)]^*. \quad (9)$$

For the equalization, a receiver at the distance of $d_{error} = d + \Delta d$ is used and a receiver at the reference distance of $d = 1\,m$ for the de-equalization. As shown in [57], this compensation almost completely eliminates impairments on spatial upsampling due to distance errors. It can be assumed that this compensation can without any restrictions be applied for voice directivities.

3 TECHNICAL EVALUATION

The following section analyzes the performance of the SUpDEq method for spatial upsampling of voice directivity. Based on the measured directivity of the HEAD acoustics HMS II.3, the reference dataset (Lebedev grid with 2,702 captured directions) is compared to sparse sets processed with SUpDEq according to Sec. 1 and datasets obtained by means of SH interpolation without pre or postprocessing in the following denoted as unprocessed. The first part of this section examines the sequential measurement procedure. The frequency-dependent directivity patterns and frequency responses for some selected directions are compared to the reference set and the averaged spectral differences are calculated. The second part examines if the results can be transferred to measurements with the SMA. The differences between the different grid types (pentakis and Lebedev grid) are analyzed, as well as differences between the sequential and simultaneous measurement procedure.

3.1 Sequential Array

First, we analyzed the directivity in the horizontal and vertical plane determined for different orders $N_{low} = 4$, 7, 10, 13 for SUpDEq-processed and unprocessed datasets. Fig. 3 ($N_{low} = 4$, 7) and Fig. 4 ($N_{low} = 10$, 13) show the patterns for four different octave bands. We refrained from plotting the directivity patterns for lower frequencies, as at the examined orders below 1 kHz no aliasing or truncation errors occur either with or without SUpDEq processing. In the octave bands of 1 kHz and 2 kHz, the plots show only slight differences between the different sparse grids and the reference. For the unprocessed sets at $N_{low} = 4$, 7 in the 4 kHz and 8 kHz octave band, the deviations increase and are spread over various directions in both the horizontal and vertical plane reaching 10 dB and more for various directions. For the SUpDEq-processed sets the deviations are significantly lower and concentrated to rearward directions. In the frontal hemisphere they do not exceed 3 dB.

For higher spatial orders of $N_{low} = 10$, 13 the deviations are generally smaller, both for the unprocessed and SUpDEq-processed sets. For frequencies up to 4 kHz they are in the frontal hemisphere below 2 dB for the SUpDEq-processed sets and do not exceed 4 dB for the unprocessed sets. This shows that SUpDEq-processing of sparse datasets leads to directivity patterns that are much closer to the reference than the unprocessed ones.

In a next step, we considered the magnitude for frontal ($\phi = 0°$, $\theta = 0°$) and a more critical lateral ($\phi = 90°$, $\theta = 0°$) sound radiation at $N_{low} = 4$. Even though the most important frequency range of the human voice is below 8 kHz, recent studies [47, 14] claim that higher frequencies also carry important information. Thus we plotted in Fig. 5 the magnitude responses for frequencies up to 16 kHz. Fig. 5(a) shows the magnitude for the frontal direction. For frequencies below 8 kHz the deviations of the SUpDEq-processed datasets reach 2.3 dB at $N_{low} = 4$ and are less than 1.8 dB at $N_{low} = 7$. The unprocessed datasets show a maximum of 13 dB at $N_{low} = 4$ and 8.1 dB at $N_{low} = 7$. As shown in Fig. 5(b), for the lateral direction the same trend can be observed, but the deviations are especially for the unprocessed datasets larger. For frequencies up to 8 kHz they reach 13 dB at 2.8 kHz for $N_{low} = 4$ and more than 20 dB for $N_{low} = 7$ at 5.5 kHz and 6.2 kHz. For the SUpDEq-method we observed a maximum of 4.3 dB at $N_{low} = 4$ and 3 dB at $N_{low} = 7$. At frequencies above 8 kHz the deviations increase further for nearly all conditions. It is only when applying the SUpDEq method at $N_{low} = 7$ that the maximal deviations remain below 6 dB. In general, as also observed in the directivity plots, frontal directions are less

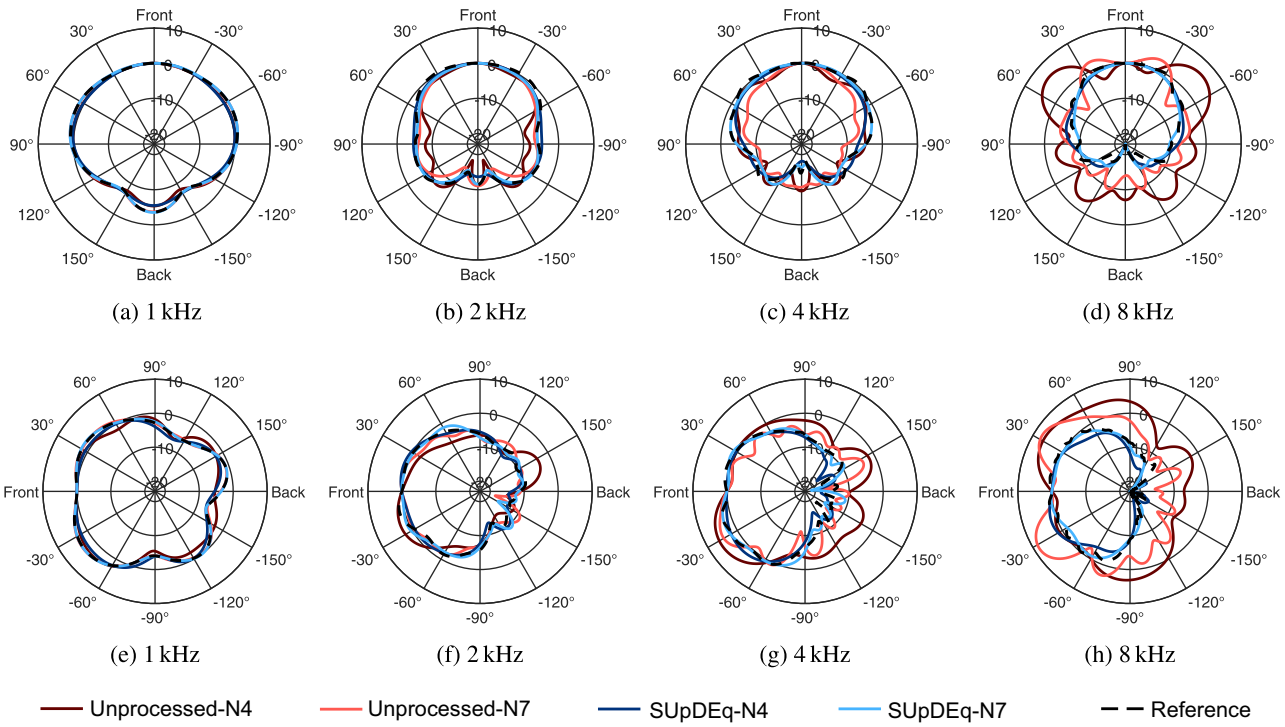


Fig. 3. Directivity in the horizontal plane (a–d) and vertical plane (e–h) determined from the reference set (black, dashed) as well as from the unprocessed (red) or SUpDEq-processed (blue) sets at $N_{low} = 4, 7$ based on the sequential measurements, normalized for frontal sound radiation. (a,e): 1 *kHz*, (b,f): 2 *kHz*, (c,g): 4 *kHz*, and (d,h): 8 *kHz* octave band.

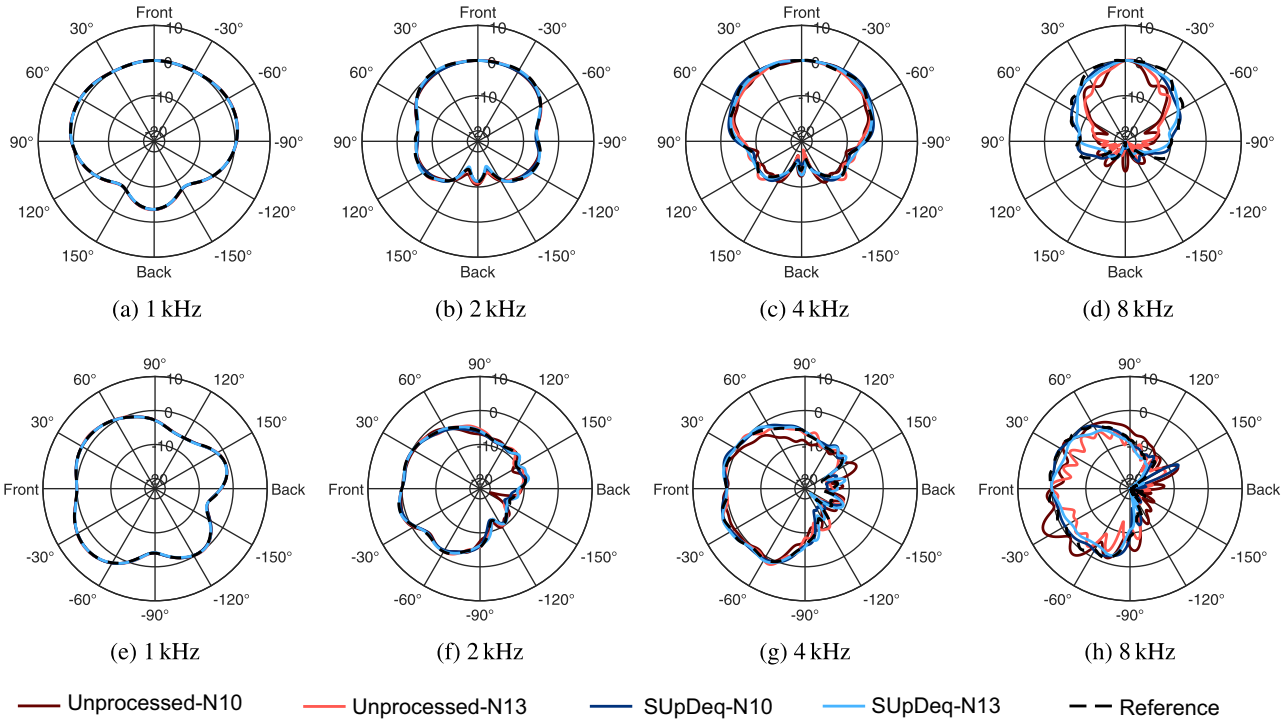
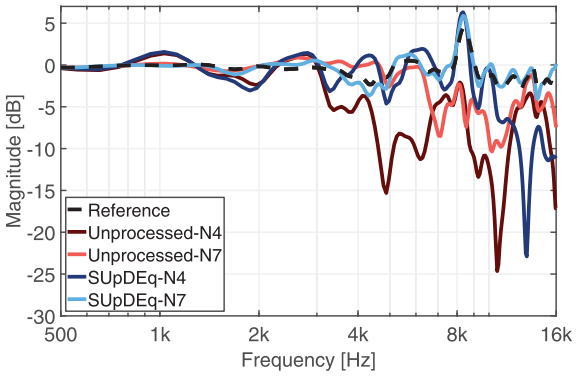
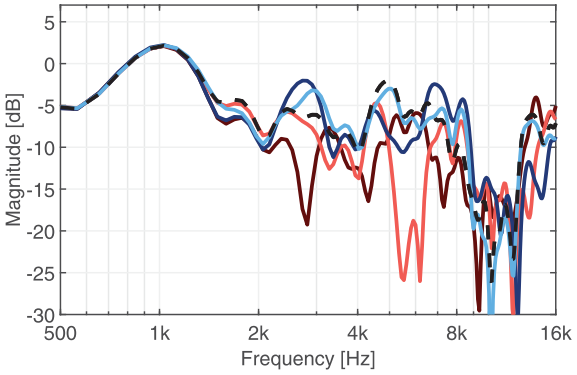


Fig. 4. Directivity in the horizontal plane (a–d) and vertical plane (e–h) determined from the reference set (black, dashed) as well as from the unprocessed (red) or SUpDEq-processed (blue) sets at $N_{low} = 10, 13$ based on the sequential measurements, normalized for frontal sound radiation. (a,e): 1 *kHz*, (b,f): 2 *kHz*, (c,g): 4 *kHz*, and (d,h): 8 *kHz* octave band.



(a) Frontal



(b) Lateral

Fig. 5. Magnitude responses for the reference (black), unprocessed (red), and SUpDEq-processed (blue) radiation. The unprocessed and SUpDEq-processed sets are based on a sparse directivity set with 38 sampled directions ($N_{low} = 4$) and 86 sampled directions ($N_{low} = 7$). (a): Frontal direction ($\phi = 0°$, $\theta = 0°$). (b): Lateral direction ($\phi = 90°$, $\theta = 0°$).

critical than lateral or rearward directions, mostly because diffraction of the head influences sound radiation strongest for these directions and is not completely matched by the applied (de-)equalization dataset. A detailed discussion of this issue that we observed in a very similar way for HRTFs can be found in [40].

Next, we analyzed the spectral deviations from the reference set as a function of N on a Lebedev grid with 2,702 sampling points as test sampling grid $\Omega_t = \{(\phi_1, \theta_1), \ldots, (\phi_T, \theta_T)\}$. For this, the frequency-dependent spectral differences per sampling point were calculated in dB as

$$\Delta G(\omega, \Omega_t) = 20lg \frac{\mid p_{dir,ref}(\omega, \Omega_t) \mid}{\mid p_{dir,test}(\omega, \Omega_t) \mid}, \tag{10}$$

where $p_{dir,ref}$ is the directivity extracted from the reference set measured on a dense grid and $p_{dir,test}$ is the directivity calculated from the unprocessed or SUpDEq-processed dataset at the directions Ω_t. Then, the absolute value of $\Delta G(\omega, \Omega_t)$ was averaged over all sampling points to obtain the frequency-dependent measure $\Delta G_f(\omega)$ (in dB):

$$\Delta G_f(\omega) = \frac{1}{n_{\Omega_t}} \sum_{\Omega_t=1}^{n_{\Omega_t}} \mid \Delta G(\omega, \Omega_t) \mid, \tag{11}$$

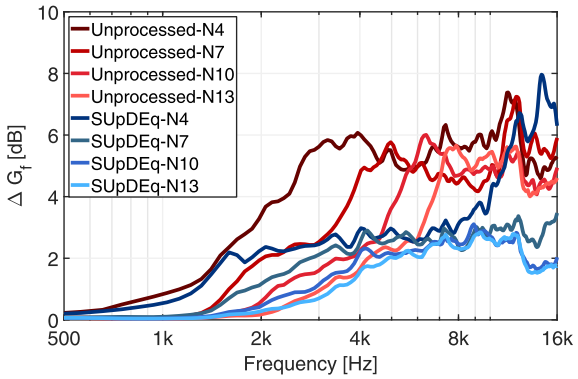


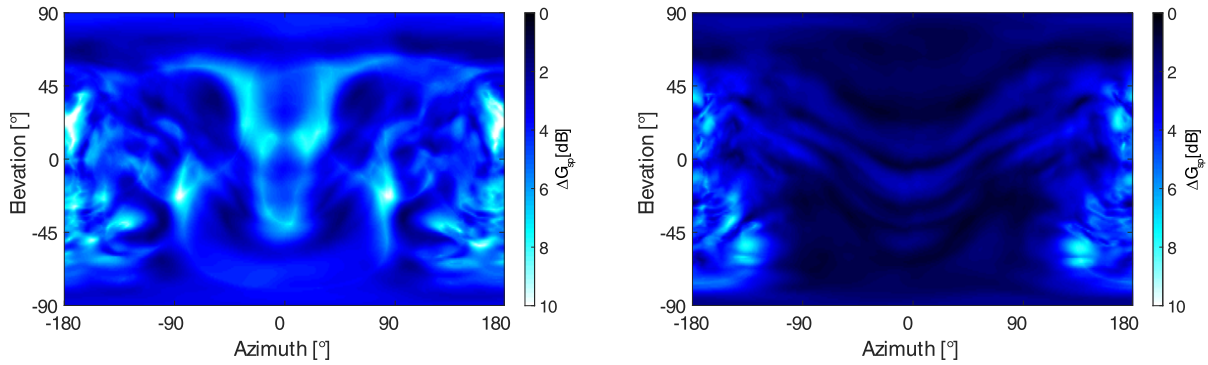
Fig. 6. Spectral differences $\Delta G_f(\omega)$ between the reference set and unprocessed (red) or SUpDEq-processed (blue) sets at $N_{low} = 4, 7, 10, 13$ based on the sequential measurements with the VariSphear device.

Fig. 6 presents the frequency-dependent spectral differences $\Delta G_f(\omega)$ at $N_{low} = 4, 7, 10, 13$. The plot clearly shows that the spectral differences are significantly smaller for the SUpDEq method than for the unprocessed datasets. For the unprocessed datasets, the spectral differences increase rapidly above 2 dB at the respective aliasing frequency between 2 kHz and 6 kHz depending on N. For the SUpDEq method, the spectral differences are generally lower and have a more gentle rise. Here, the differences exceed 2.5 dB at frequencies between 3 kHz and 8 kHz and remain below 4 dB up to 10 kHz but increase sharply for $N_{low} = 4$ above 10 kHz. The curve has a similar shape to that of the unprocessed dataset, only shifted toward higher frequencies. The sharp increase most likely represents spatial aliasing artifacts caused by the SFT of the equalized datasets. For higher spatial orders $N_{low} = 7, 10, 13$, the spectral differences remain below 4 dB up to 16 kHz.

An analysis of the spatial distributions of the deviations concludes the spectral analysis. For this, we calculated the absolute value of $\Delta G(\omega, \Omega_t)$ averaged across the angular frequency ω to obtain one value $\Delta G_{sp}(\Omega_t)$ (in dB) per sampling point:

$$\Delta G_{sp}(\Omega_t) = \frac{1}{n_\omega} \sum_{\omega=1}^{n_\omega} \mid \Delta G(\omega, \Omega_t) \mid . \tag{12}$$

Fig. 7 shows the spectral differences $\Delta G_{sp}(\Omega_t)$ per sampling point for the unprocessed and SUpDEq-processed datasets at $N_{low} = 4$. As most of the energy of speech is below 8 kHz, we averaged the spectral differences for $f \leq 8$ kHz. In this case, the test sampling grid Ω_t was full spherical calculated for ϕ and θ in steps of 1°. The plots show that independent of the method for spatial upsampling, the spectral differences are maximal for directions to the rear. Spectral differences are generally higher for the unprocessed datasets, with a maximum of $\Delta G_{sp}(\Omega_t) = 12.8$ dB at $\phi = -171°$ and $\theta = 26°$. The SUpDEq method results in a maximum spectral difference $\Delta G_{sp}(\Omega_t)$ of 8.3 dB at $\phi = -148°$ and $\theta = -39°$.



(a) Unprocessed (b) SUpDEq

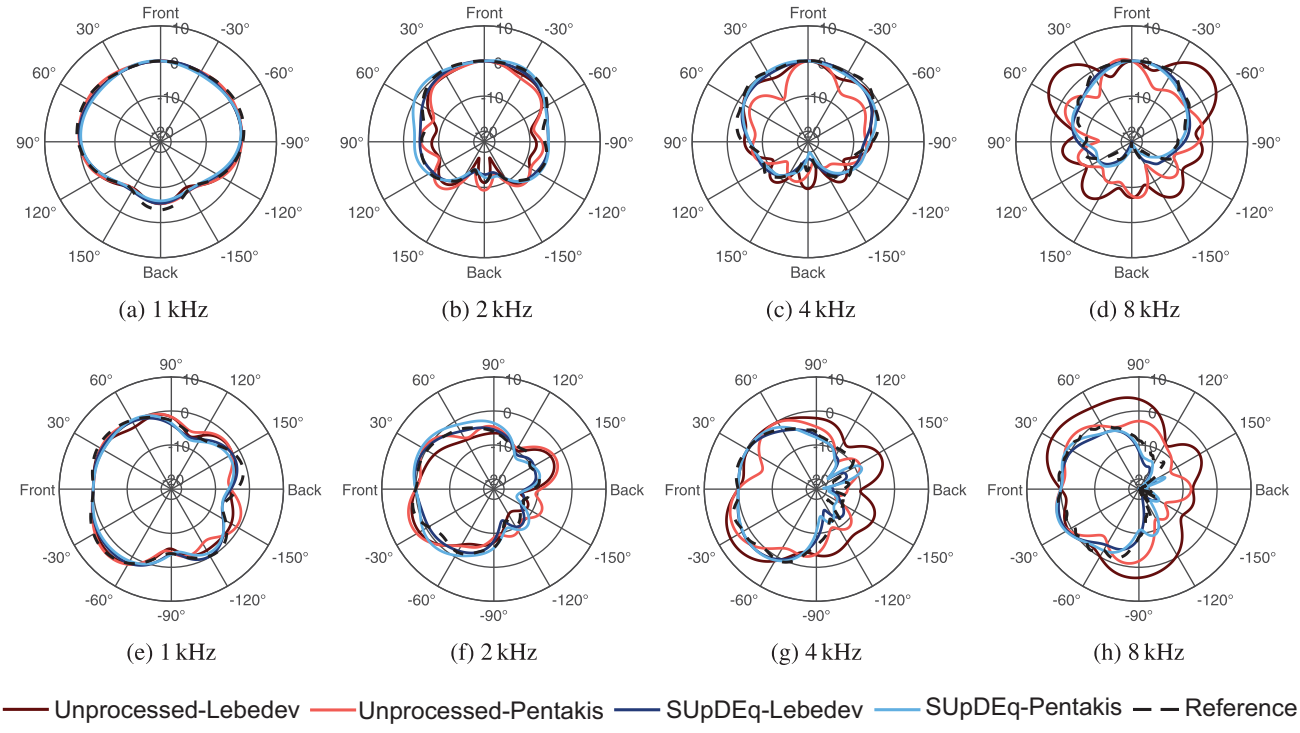
Fig. 7. Spectral differences $\Delta G_{sp}(\Omega_t)$ between the reference set and unprocessed (a) or SUpDEq-processed (b) set for sequential measurements with the VariSphear device, at $N_{low} = 4$ and $f \leq 8$ kHz.

3.2 Surrounding Microphone Array

This section examines whether the spatial upsampling of data measured simultaneously with the SMA (see Sec. 2.2) performs comparably to the spatial upsampling of data measured sequentially with the VariSphear on a Lebedev grid satisfying $N_{low} = 4$. The analysis is split up into two parts. The first part analyzes whether the sampling scheme has a significant influence on the performance of the SUpDEq method. The second part then examines whether the different setups measuring the directivity in a sequential or simultaneous procedure affect the results of the spatial upsampling.

3.2.1 *Influence of Sampling Scheme*

To analyze the effects of the sampling scheme, we again used the sequential measurements performed with the VariSphear. Fig. 8 depicts the directivity determined from the SUpDEq-processed and unprocessed datasets at different octave bands for the Lebedev and pentakis grids. While the unprocessed directivity patterns show some differences compared to the reference, especially toward higher frequencies, the SUpDEq-processed patterns are quite similar and match the reference quite well. Only for rearward directions, larger differences occur. Fig. 10 presents the frequency-dependent spectral differences $\Delta G_f(\omega)$ for the unprocessed and SUpDEq-processed datasets at $N_{low} = 4$



(a) 1 kHz (b) 2 kHz (c) 4 kHz (d) 8 kHz

(e) 1 kHz (f) 2 kHz (g) 4 kHz (h) 8 kHz

Fig. 8. Directivity in the horizontal plane (a–d) and vertical plane (e–h) determined from the reference set (black, dashed) as well as from the unprocessed (red) or SUpDEq-processed (blue) set applying sequential measurements on the pentakis grid and the Lebedev grid at $N_{low} = 4$. (a,e): 1 *kHz*, (b,f): 2 *kHz*, (c,g): 4 *kHz*, and (d,h): 8 *kHz* octave band, normalized for frontal sound radiation.

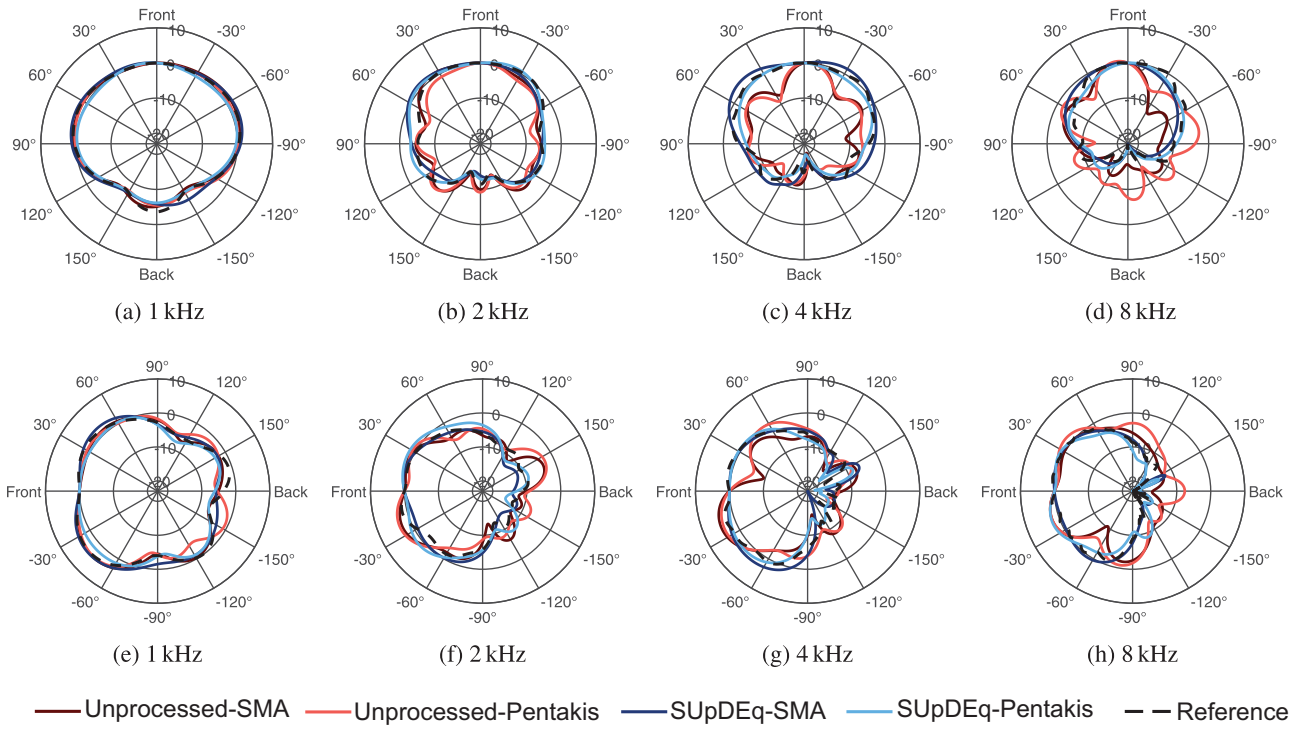


Fig. 9. Directivity in the horizontal plane (a–d) and vertical plane (e–h) determined from the reference set (black, dashed) as well as from the unprocessed (red) or SUpDEq-processed (blue) set applying a simultaneous measurement with the SMA and a sequential measurement performed with the VariSphear, both on the pentakis grid at $N_{low} = 4$, normalized for frontal sound radiation. (a,e): 1 *kHz*, (b,f): 2 *kHz*, (c,g): 4 *kHz*, and (d,h): 8 *kHz* octave band.

and with respect to the sampling scheme (pentakis or Lebedev grid). Some small differences can be observed for the unprocessed sets, mainly between 2 kHz and 5 kHz. Here, the pentakis grid performs slightly better. For the SUpDEq-processed sets, the differences to the reference are about the same, i.e., the grid-type has only a very small influence. The evaluation thus shows that the pentakis grid with its 32 sampling points performs similar to a Lebedev grid that fulfills $N_{low} = 4$ (38 sampling points) and each of these grids can be applied to measure voice directivity for frequency bands up to 8 kHz.

3.2.2 Influence of Measurement Procedure

Finally, we analyzed the influence of the specific measurement procedure and compared the results of the sequential to the simultaneous measurements. Fig. 9 depicts the directivity patterns determined with the SMA (parallel measurement) and compares them to measurements on the same pentakis grid with the VariSphear device (sequential measurement). Especially for the SUpDEq-processed datasets, there are only small differences between the directivity patterns. Analyzing the spectral differences $\Delta G_f(\omega)$ for both array types at $N_{low} = 4$ (see Fig. 10) confirms that the measurement procedures perform comparably. For $f \leq$ 8 kHz the spectral deviations differ less than 1 dB for the SUpDEq-processed datasets and 2 dB for the unprocessed datasets. While the deviations for all grids remain below 4 dB up to 8 kHz, for higher frequencies spectral deviations of up to 8 dB can be observed, making a meaningful use of the datasets rather difficult here.

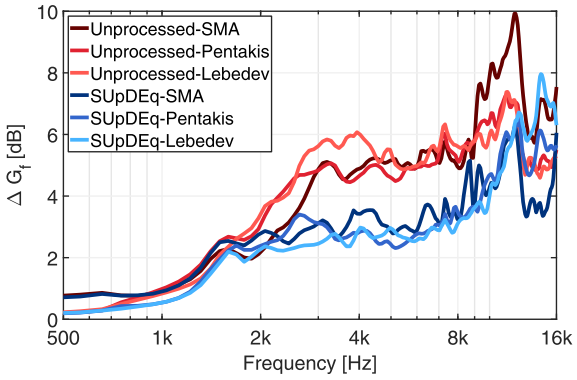
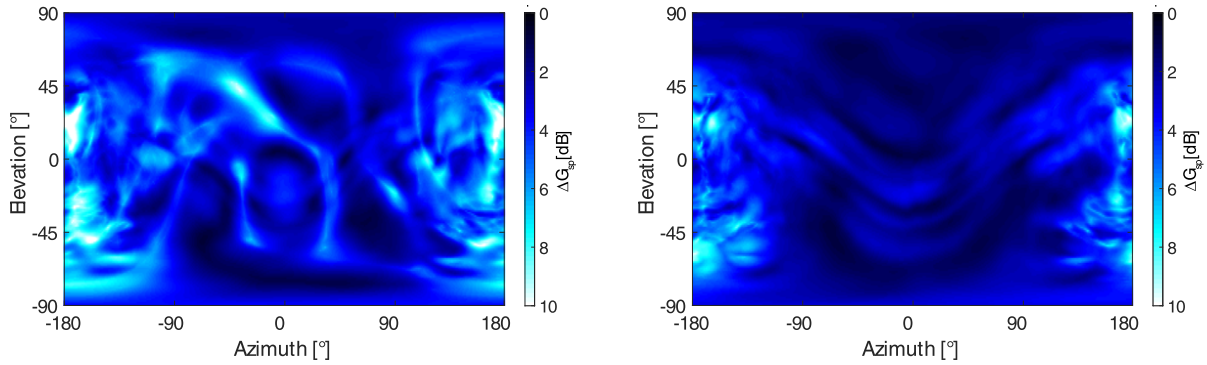


Fig. 10. Spectral differences $\Delta G_f(\omega)$ between the reference set and unprocessed (red) or SUpDEq-processed (blue) set for the SMA (simultaneous measurement), the pentakis grid (sequential measurement), and the Lebedev grid (sequential measurement) at $N_{low} = 4$.

Fig. 11 shows the spectral differences $\Delta G_{sp}(\Omega_t)$ per sampling point between the reference set and the unprocessed or SUpDEq-processed directivity set at $N_{low} = 4$ and $f \leq 8$ kHz, originally captured with the SMA. The spectral differences of the simultaneous measurements with the SMA are quite similar to the differences of the sequential measurements performed with the VariSphear (cf. Fig. 7). In detail, the results vary a little, as the SMA measurements show slightly higher maximal deviations than the VariSphear measurements, resulting in $\Delta G_{sp}(\Omega_t) = 13.5$ dB at $\phi = -171°$ and $\theta = 26°$ for the unprocessed set and $\Delta G_{sp}(\Omega_t) =$



(a) Unprocessed (b) SUpDEq

Fig. 11. Spectral differences $\Delta G_{sp}(\Omega_t)$ between the reference set and unprocessed (a) or SUpDEq-processed (b) set at $N_{low} = 4$ and $f \leq 8$ kHz for the SMA (simultaneous measurement).

Table 1. Spectral differences ΔG in dB between the reference set and unprocessed or SUpDEq-processed sets for the different sampling schemes and measurement procedures at $f \leq 8$ kHz.

Sampling Scheme / Procedure	Unproc.	SUpDEq
SMA-N4	4.00	2.75
Pentakis-N4	4.00	2.40
Lebedev-N4	4.34	2.27
Lebedev-N7	3.32	1.91
Lebedev-N10	2.73	1.52
Lebedev-N13	2.01	1.33

10.0 dB at $\phi = 176°$ and $\theta = 26°$ for the SUpDEq-processed set.

Finally, we compare the different datasets based on a single value. For this, we average the spectral differences across ω and Ω_t:

$$\Delta G = \frac{1}{n_{\Omega_t}} \frac{1}{n_\omega} \sum_{\Omega_t=1}^{n_{\Omega_t}} \sum_{\omega=1}^{n_\omega} \mid \Delta G(\omega, \Omega_t) \mid . \tag{13}$$

Table 1 shows ΔG for all sampling schemes and measurement procedures investigated in this study. The SUpDEq method reduces ΔG for all datasets by at least one third (in dB). Regarding the SUpDEq-processed sets, ΔG is 0.35 dB larger for the SMA than the for the pentakis grid. As both are based on the same sampling scheme, this difference might be a result of positioning inaccuracies in the SMA. However, it can be concluded that measurements with the SMA perform similarly to measurements with the VariSphear on the same grid. ΔG varies less than 0.5 dB at $N_{low} = 4$ between the procedures and setups both for the unprocessed and SUpDEq-processed datasets.

4 DISCUSSION

4.1 Sampling Scheme and Measurement Procedure

The evaluation revealed that the unprocessed datasets show distinct spectral differences to a reference spread over the entire angular range. In contrast, the SUpDEq method performs much better and only shows small deviations from the reference for frontal or lateral directions. The spatial equalization before the SFT reduces the energy in the higher spatial orders, and by this the influence of spatial aliasing and truncation errors in the SH-transformed datasets decreases. This results in an improved upsampling and lower reconstruction errors when applying the inverse SFT. Greater differences only remain for sound waves radiated to the rearward directions. The propagation of sound around the head or a rigid sphere leads to a bright spot at the rearward direction because the waves propagating around the sphere are in phase there and thus interfere constructively [59]. However, this interference pattern changes rapidly for adjacent directions, especially toward higher frequencies. As the equalization function does not exactly match this interference pattern, energy still remains at higher spatial orders after equalization. A similar behavior can be observed for HRTFs at contralateral directions and has been discussed in greater detail in [40].

We showed in a recent study [60] that the specific sampling scheme (grid type) has only a minor influence on the results. In the present study, we further confirmed this by showing that at $N_{low} = 4$, both the Lebedev grid with 38 sampling points and pentakis grid with 32 sampling points perform comparably. The different procedures of sequential measurements with the VariSphear device or simultaneous measurements with the SMA have only a minor influence on the acquired directivity and average spectral differences to the reference resulting in an increase of the averaged spectral differences by 0.35 dB for the SUpDEq-processed dataset. This small increase might be caused by remaining influences of positioning inaccuracies of the dummy head in the SMA that were not corrected by the distance error compensation (see Sec. 2.5).

Our results in Figs. 6 and 10 reveal that above 8 kHz the spectral errors become large and exceed 6 dB at $N_{low} = 4$. Thus, applying the SUpDEq method to SMA measurements with 32 microphones allows us to efficiently determine the directivity of a head and mouth simulator for frequency bands up to 8 kHz and might also be appropriate to determine voice directivity of human speakers in this frequency range. A direct comparison of the presented voice directivity measurements to other studies is difficult,

as only a few of them deal with dummy heads with integrated mouth simulators, e.g., [7, 23, 8, 9], and none of them examined a HEAD acoustics dummy head. However, the general shape of the captured radiation pattern is comparable to the measurements of a B&K dummy head in the horizontal and vertical plane [7] and for spherical datasets [23]. However, these studies do not consider or discuss any methods for spatial upsampling of sparse datasets. Apart from Chu and Warnock [23], who published average values and standard deviations in third-octave bands for measurements on 40 subjects, no spherical data of dummy heads or human speakers are available, and thus it is hardly possible to compare our spherically measured directivity to other studies. This underlines the need for suitable methods to capture and spatially upsample spherical radiation patterns of voice directivity.

4.2 Application to Human Voice Directivity

In this study we evaluated the SUpDEq method based on a dummy head and mouth simulator because no high-density datasets of human voice directivity are available that could serve as a reference. However, when applying the proposed methods to human voice directivity, there are several issues that additionally need to be considered.

4.2.1 Radiation Characteristics

The SUpDEq method relies on the assumption that human voice sound radiation can be approximated by a point source positioned at the mouth opening, which is suitable for a dummy head and mouth simulator but needs to be proven for a human speaker. In this context, the influence of glottis and nose on sound radiation and toward lower frequencies the radiation from the human skull might have an impact. However, it has already been shown by Dunn and Farnsworth [5] that for the human voice mainly the radiation from the mouth opening contributes to the radiation in the far-field. Halkosaari et al. [8] found relevant differences between a dummy head and human voice directivity, which are mainly caused by the size difference of the mouth between the dummy head and average test subject. However, their study showed as well that the directivity of a dummy head reflects typical spatial properties of a human voice. Accordingly, we can assume that the methods of spatial upsampling as applied in this paper perform comparably for dummy head and human voice excitation.

4.2.2 Voice Excitation Signal

All datasets used in this study are based on impulse response measurements with sine sweep excitation, which is common, e.g., for loudspeaker measurements. However, the results of our investigations can be directly transferred to human voice signals, because the complete processing chain of SUpDEq only consists of linear functions and operations, and it is thus irrelevant if the processing is applied to impulse responses or any other kind of signals, e.g., to fluent speech of a human talker. Alternatively, when investigating the directivity pattern of single phonemes, impulse responses can be determined for any specific direction related to the frontal direction from the voice signal. In this case, for vocals the glissando method [22] can be applied to obtain appropriate excitation signals, and for fricatives the voice signal can be regarded as spectrally shaped noise. Of course, for the measured signals an appropriate signal-to-noise ratio in each considered frequency band needs to be ensured, which relates to the self-noise of the microphones and preamplifiers and to the amplitude differences between human voice excitation and background signals.

4.2.3 Low-Frequency Effects

Toward lower frequencies, the measurements involve further limitations and are affected by reflections and room modes of the measuring room, which in our study are relevant below 200 Hz and thus have an influence around the human fundamental frequency. For HRTFs a well-suited approach to compensate for these effects is to replace the low-frequency range of the measured transfer function by an analytical expression (e.g., [52]). According to the principle of reciprocity [26] this approach can without any restrictions be adapted to directivity measurements. Thus, as a further postprocessing step, a low-frequency extension, e.g., in the frequency domain according to [61], could be performed, which substitutes the original low-frequency component by an adequately matched one.

4.2.4 Time Variances

Time variances due to movements of a human speaker inside the SMA during the measurement are a further issue that can influence the spatial upsampling. To equalize these time-variant off-center positions of the head, the distance error compensation, which we already integrated in the processing of the measured signals (Sec. 2.5), needs to work adaptively over time. For this, the directivity has to be determined separately for small audio frames, e.g., with a length of 20 ms for fluent speech. In this case, the complete SUpDEq processing as proposed in Sec. 1 remains unchanged, and it only has to be ensured that movements of the human speaker within each time frame are so small that a quasi-static processing is suitable. Furthermore, such a frame-based processing allows resolving time variances of voice directivity of fluent speech, which is an important issue to be investigated in future studies.

5 CONCLUSION

We presented an SH-based approach for spatial upsampling (interpolation) of voice directivity measured on a sparse spherical grid and for this adopted the SUpDEq method, which was originally designed for spatial upsampling of sparse HRTF sets. We showed that similar to HRTFs, a rigid sphere transfer function can be applied for spatial equalization of measured directivity. The equalization transforms energy from high spatial orders to lower orders. As a result, spatial aliasing and order-truncation errors, which are caused by the spatial upsampling, are significantly reduced. Finally, the de-equalization recovers energy at higher spatial orders. The evaluation revealed that

the SUpDEq-processed datasets lead to directivity patterns that are much closer to a reference set than the unprocessed datasets. Accordingly, for a dummy head equipped with a mouth simulator, already at $N_{low} = 4$ a decent full-spherical dense directivity set can be generated for frequencies up to 8 kHz. The results show that a surrounding microphone array with a number of 32 microphones can be used to reliably determine voice directivity.

The study forms a basis for human voice directivity measurements. Applying the SUpDEq method to directivity measurements made with a surrounding microphone array allows loudness and articulation-related influences to be analyzed with a practically feasible experimental setup. This reduces many of the difficulties and inaccuracies caused by a low spatial resolution. By this, limitations of many studies can be overcome, as up to now only the horizontal or vertical plane was examined most of the time because of the required high number of measurement directions. The proposed method can even be applied to data obtained in previous measurements (e.g., [13]) to improve the interpolation of sparsely measured human voice directivity.

In the context of this study, it is of great relevance to analyze the perceptual influence of higher spatial orders of a voice directivity. Thus, psychoacoustic experiments need to be performed in order to analyze the minimum required spatial order. Within this scope, a first approach to perceptually evaluate the required order of a source directivity has been recently presented [62]. Our study and the described methods to determine (individual) human voice directivity form an important basis for such investigations.

6 ACKNOWLEDGMENT

The authors would like to thank Raphael Gillioz and Melissa Ramírez for supporting the measurements. The research presented in this paper has been carried out in the Research Project NarDasS that was funded by the Federal Ministry of Education and Research in Germany, support code: BMBF 03FH014IX5-NarDasS. A Matlab-based implementation of the SUpDEq method is available on github.com/AudioGroupCologne/SUpDEq.

7 REFERENCES

[1] G. Saunders, *Treatise on Theaters* (I. and J. Taylor, London, 1790).

[2] B. Wyatt, *Observation on the Design for the Theatre Royal, Drury Lane* (J. Taylor, London, 1813).

[3] J. Henry, "Annual Report of the Board of Regents of the Smithsonian Institution," *Tech. rep.*, A. G. F. Nicholson, Washington, DC (1857).

[4] F. Trendelenburg, "Beitrag zur Frage der Stimmrichtwirkung," *Zeitschrift für techn. Physik*, vol. 11, pp. 558–563 (1929).

[5] H. K. Dunn and D. W. Farnsworth, "Exploration of Pressure Field Around the Human Head During Speech," *J. Acoust. Soc. Am.*, vol. 10, pp. 184–199 (1939), doi:https://doi.org/10.1121/1.1915975.

[6] J. L. Flanagan, "Analog Measurements of Sound Radiation From the Mouth," *J. Acoust. Soc. Am.*, vol. 32, no. 12, pp. 1613–1620 (1960), doi:https://doi.org/10.1121/1.1936423.

[7] J. Huopaniemi, K. Kettunen, and J. Rahkonen, "Measurement and Modeling Techniques for Directional Sound Radiation From the Mouth," *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pp. 183–186 (1999), doi:https://doi.org/10.1109/ASPAA.1999.810880.

[8] T. Halkosaari, M. Vaalgamaa, and M. Karjalainen, "Directivity of Artificial and Human Speech," *J. Audio Eng. Soc.*, vol. 53, no. 7/8, pp. 620–631 (2005 Jul.).

[9] G. Fischer, C. Schneiderwind, and A. Neidhardt, "Comparing the Directivity of a Mouth Simulator and a Simple Physical Model," presented at the *Proceedings of the 45th DAGA* (2019).

[10] C. Pörschmann, "One's Own Voice in Auditory Virtual Environments," *Acta Acust. united Acust.*, vol. 87, no. 3, pp. 378–388 (2001).

[11] J. M. Arend, T. Lübeck, and C. Pörschmann, "A Reactive Virtual Acoustic Environment for Interactive Immersive Audio," presented at the *2019 AES International Conference on Immersive and Interactive Audio* (2019 Mar.), conference paper 9.

[12] A. Neidhardt, "Detection of a Nearby Wall in a Virtual Echolocation Scenario Based on Measured and Simulated OBRIRS," presented at the *2018 AES International Conference on Spatial Reproduction - Aesthetics and Science* (2018 Jul.), conference paper P1-4.

[13] M. Brandner, M. Frank, and D. Rudrich, "DirPat–Database and Viewer of 2D/3D Directivity Patterns of Sound Sources and ReceiversDatabase and Viewer of 2D/3D Directivity Patterns of Sound Sources and Receivers," presented at the *144th Convention of the Audio Engineering Society*, pp. 1–5 (2018 May), convention paper 425.

[14] P. Kocon and B. B. Monson, "Horizontal Directivity Patterns Differ Between Vowels Extracted From Running Speech," *J. Acoust. Soc. Am.*, vol. 144, no. 1, pp. EL7–EL12 (2018), doi:https://doi.org/10.1121/1.5044508.

[15] B. Katz and C. D'Alessandro, "Directivity Measurements of the Singing Voice," presented at the *Proceedings of the 19th International Congress on Acoustics* (2007).

[16] L. M. Wang and M. C. Vigeant, "Evaluations of Output From Room Acoustic Computer Modeling and Auralization Due to Different Sound Source Directionalities," *Appl. Acoust.*, vol. 69, pp. 1281–1293 (2008), doi:https://doi.org/10.1016/j.apacoust.2007.09.004.

[17] B. N. J. Postma and B. F. Katz, "Dynamic Voice Directivity in Room Acoustic Auralizations," *Proc. 42nd DAGA*, pp. 352–355 (2016).

[18] B. N. J. Postma, H. Demontis, and B. F. G. Katz, "Subjective Evaluation of Dynamic Voice Directivity for Auralizations," *Acta Acust. united Acust.*, vol. 103, no. 2, pp. 181–184 (2017).

[19] B. B. Monson, E. J. Hunter, and B. H. Story, "Horizontal Directivity of Low- and High-Frequency Energy in Speech and Singing," *J. Acoust. Soc. Am.*, vol.

132, no. 1, pp. 433–441 (2012), doi:https://doi.org/10.1121/1.4725963.

[20] A. H. Marshall and J. Meyer, "The Directivity and Auditory Impressions of Singers," *Acustica*, vol. 58, pp. 130–140 (1985).

[21] A. Moreno and J. Pfretzschner, "Human Head Directivity in Speech Emission: A New Approach," *Acoust. Lett.*, vol. 1, pp. 78–84 (1978).

[22] M. Kob and H. Jers, "Directivity Measurement of a Singer," *J. Acoust. Soc. Am.*, vol. 105, p. 1003 (1999), doi:https://doi.org/10.1121/1.425813.

[23] W. T. Chu and A. C. C. Warnock, "Detailed Directivity of Sound Fields Around Human Talkers," Tech. rep., National Research Council of Canada (2002), doi:https://doi.org/10.4224/20378930.

[24] M. Pollow, *Directivity Patterns for Room Acoustical Measurements and Simulations* (Logos Verlag Berlin, 2015).

[25] J. M. Arend, P. Stade, and C. Pörschmann, "Binaural Reproduction of Self-Generated Sound in Virtual Acoustic Environments," *Proc. 173rd Meeting Acoust. Soc. Am.*, pp. 1–13 (2017), doi:https://doi.org/10.1121/2.0000574.

[26] H. Wallach, "On Sound Localization," *J. Acoust. Soc. Am.*, vol. 10, no. 1939, pp. 270–274 (1939), doi:https://doi.org/10.1121/1.1915985.

[27] M. Pollow, K. -V. V. Nguyen, O. Warusfel, T. Carpentier, M. Müller-Trapet, M. Vorländer, and M. Noisternig, "Calculation of Head-Related Transfer Functions for Arbitrary Field Points Using Spherical Harmonics Decomposition," *Acta Acust. united Acust.*, vol. 98, no. 1, pp. 72–82 (2012), doi:https://doi.org/10.3813/AAA.918493.

[28] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and Range Extrapolation of HRTFs," *Proc. ICASSP 2004*, vol. 5, pp. 45–48 (2004), doi:https://doi.org/10.1109/ICASSP.2004.1326759.

[29] E. Wenzel and S. Foster, "Perceptual Consequences of Interpolating Head-Related Transfer Functions During Spatial Synthesis," *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pp. 102–105 (1993), doi:https://doi.org/10.1109/ASPAA.1993.379986.

[30] K. Hartung, J. Braasch, and S. J. Sterbing, "Comparison of Different Methods for the Interpolation of Head-Related Transfer Functions," presented at the *AES 16th International Conference: Spatial Sound Reproduction*, pp. 319–329 (1999 Mar.), conference paper 16-028.

[31] 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, no. 1, pp. 528–537 (2000), doi:https://doi.org/10.1121/1.428321.

[32] E. G. Williams, *Fourier Acoustics - Sound Radiation and Nearfield Acoustical Holography* (Academic Press, London, UK, 1999).

[33] B. Rafaely, *Fundamentals of Spherical Array Processing* (Springer-Verlag, Berlin Heidelberg, 2015).

[34] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Analyzing Head-Related Transfer Function Measurements Using Surface Spherical Harmonics," *J. Acoust. Soc. Am.*, vol. 104, no. 4, pp. 2400–2411 (1998), doi:https://doi.org/10.1121/1.423749.

[35] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral Equalization in Binaural Signals Represented by Order-Truncated Spherical Harmonics," *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4087–4096 (2017), doi:https://doi.org/10.1121/1.4983652.

[36] D. L. Alon, Z. Ben-Hur, B. Rafaely, and R. Mehra, "Sparse Head-Related Transfer Function Representation With Spatial Aliasing Cancellation," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 6792–6796 (2018), doi:https://doi.org/10.1109/ICASSP.2018.8462101.

[37] C. Pike and A. Tew, "Subjective Assessment of HRTF Interpolation With Spherical Harmonics," presented at the *Proceedings of the International Conference on Spatial Audio - ICSA 2017* (2017).

[38] M. Zaunschirm, C. Schoerkhuber, and R. Hoeldrich, "Binaural Rendering of Ambisonic Signals by HRIR Time Alignment and a Diffuseness Constraint," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3616–3627 (2018), doi:https://doi.org/10.1121/1.5040489.

[39] F. Brinkmann and S. Weinzierl, "Comparison of Head-Related Transfer Functions Pre-Processing Techniques for Spherical Harmonics Decomposition," presented at the *2018 AES International Conference on Audio for Virtual and Augmented Reality*, pp. 1–10 (2018 Aug.), conference paper P9-3.

[40] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 6, pp. 1060–1071 (2019), doi:https://doi.org/10.1109/TASLP.2019.2908057.

[41] C. Pörschmann and J. M. Arend, "A Method for Spatial Upsampling of Directivity Patterns of Human Speakers by Directional Equalization," *Proc. 45th DAGA*, pp. 1458–1461 (2019).

[42] Z. Ben-Hur, D. L. Alon, B. Rafaely, and R. Mehra, "Loudness Stability of Binaural Sound With Spherical Harmonic Representation of Sparse Head-Related Transfer Functions," *EURASIP J. Audio Speech Music Process.*, vol. 5 (2019), doi:https://doi.org/10.1186/s13636-019-0148-x.

[43] N. R. Shabtai, G. Behler, M. Vorländer, and S. Weinzierl, "Generation and Analysis of an Acoustic Radiation Pattern Database for Forty-One Musical Instruments," *J. Acoust. Soc. Am.*, vol. 141, no. 2, pp. 1246–1256 (2017), doi:https://doi.org/10.1121/1.4976071.

[44] B. Rafaely, "Analysis and Design of Spherical Microphone Arrays," *IEEE Trans. Speech Audio Proc.*, vol. 13, no. 1, pp. 135–143 (2005), doi:https://doi.org/10.1109/TSA.2004.839244.

[45] B. Bernschütz, A. Vázquez Giner, C. Pörschmann, and J. M. Arend, "Binaural Reproduction of Plane Waves With Reduced Modal Order," *Acta Acust. united Acust.*, vol. 100, no. 5, pp. 972–983 (2014), doi:https://doi.org/10.3813/AAA.918777.

[46] R. V. L. Hartley and T. C. Fry, "The Binaural Location of Pure Tones," *Phys. Rev.*, vol. 18, no. 6, pp. 431–442 (1921), doi:https://doi.org/10.1103/PhysRev.18.431.

[47] R. Blandin, A. V. Hirtum, X. Pelorson, and R. Laboissière, "The Effect on Vowel Directivity Patterns of Higher Order Propagation Modes," *J. Sound Vibr.*, vol. 432, pp. 621–632 (2018), doi:https://doi.org/10.1016/j.jsv.2018.06.053.

[48] W. Zhang, T. D. Abhayapala, R. A. Kennedy, and R. Duraiswami, "Insights Into Head-Related Transfer Function: Spatial Dimensionality and Continuous Representation," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2347–2357 (2010), doi:https://doi.org/10.1121/1.3336399.

[49] 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, no. 12, pp. 2249–2262 (2019), doi:https://doi.org/10.1109/TASLP.2019.2945479.

[50] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," *Proc. 44th DAGA*, pp. 339–342 (2018).

[51] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a Spherical-Head Model From Anthropometry," *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 472–479 (2001 Jun.).

[52] B. Bernschütz, "A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100," *Proc. 39th DAGA*, pp. 592–595 (2013).

[53] B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl, "SOFiA - Sound Field Analysis Toolbox," *Proc. Int. Conf. Spatial Audio - ICSA*, pp. 8–16 (2011).

[54] F. Brinkmann and S. Weinzierl, "AKtools–An Open Software Toolbox for Signal Acquisition, Processing, and Inspection in Acoustics," presented at the *142nd Convention of the Audio Engineering Society*, pp. 1–6 (2017 May), convention paper 309.

[55] B. Bernschütz, C. Pörschmann, S. Spors, and S. Weinzierl, "Entwurf und Aufbau eines variablen sphärischen Mikrofonarrays für Forschungsanwendungen in Raumakustik und Virtual Audio," *Proc. 36th DAGA*, pp. 717–718 (2010).

[56] V. I. Lebedev, "Spherical Quadrature Formulas Exact to Orders 25–29," *Siberian Math. J.*, vol. 18, no. 1, pp. 99–107 (1977), doi:https://doi.org/10.1007/BF00966954.

[57] C. Pörschmann and J. M. Arend, "How Positioning Inaccuracies Influence the Spatial Upsampling of Sparse Head-Related Transfer Function Sets," *Proc. Int. Conf. Spatial Audio - ICSA 2019*, pp. 1–8 (2019).

[58] J. M. Arend and C. Pörschmann, "Synthesis of Near-Field HRTFs by Directional Equalization of Far-Field Datasets," *Proc. 45th DAGA*, pp. 1454–1457 (2019).

[59] R. O. Duda and W. L. Martens, "Range Dependence of the Response of a Spherical Head Model," *J. Acoust. Soc. Am.*, vol. 104, no. 5, pp. 3048–3058 (1998), doi:https://doi.org/10.1121/1.423886.

[60] J. M. Arend and C. Pörschmann, "Spatial Upsampling of Sparse Head-Related Transfer Function Sets by Directional Equalization - Influence of the Spherical Sampling Scheme," presented at the *Proceedings of the 23rd International Congress on Acoustics* (2019), doi:https://doi.org/10.18154/RWTH-CONV-238939.

[61] B. Xie, "On the Low Frequency Characteristics of Head-Related Transfer Function," *Chinese J. Acoust.*, vol. 28, pp. 1–13 (2009).

[62] M. Frank and M. Brandner, "Perceptual Evaluation of Spatial Resolution in Directivity Patterns," *Proc. 45th DAGA*, pp. 74–77 (2019).

THE AUTHORS

Christoph Pörschmann

Johannes M. Arend

Christoph Pörschmann studied Electrical Engineering at the Ruhr-Universität Bochum (Germany) and Uppsala Universitet (Sweden). In 2001 he obtained his Doctoral Degree (Dr.-Ing.) from the Electrical Engineering and Information Technology Faculty of the Ruhr-Universität Bochum as a result of his research at the Institute of Communication Acoustics. Since 2004 he is Professor of Acoustics at TH Köln (Germany). His research interests are in the field of virtual acoustics, spatial hearing, and the related perceptual processes.

•

Johannes M. Arend received a B.Eng. degree in media technology from HS Düsseldorf (Germany) in 2011 and an M.Sc. degree in media technology from TH Köln (Germany) in 2014. Since 2015, he has been a Research Fellow and working toward a Ph.D. degree at TH Köln and TU Berlin in the field of spatial audio with a focus on binaural technology, auditory perception, and audio signal processing.

List of Publications

This thesis is a cumulative dissertation and includes the following publications, in which the author substantially contributed to the planning, execution, and writing.

Peer-Reviewed Journal Publications

Lübeck, T., **Arend, J. M.**, & Pörschmann, C. (2022). Binaural reproduction of dummy head and spherical microphone array data – A perceptual study on the minimum required spatial resolution. *J. Acoust. Soc. Am.*, *151*(1), 467–483. https://doi.org/10.1121/10.0009277

Arend*, J. M., Ramírez*, M., Liesefeld, H. R., & Pörschmann, C. (2021). Do near-field cues enhance the plausibility of non-individual binaural rendering in a dynamic multimodal virtual acoustic scene? *Acta Acust.*, *5*(55), 1–14. (*equal contributions). https://doi.org/10.1051/aacus/2021048

Arend*, J. M., Lübeck*, T., & Pörschmann*, C. (2021). Efficient binaural rendering of spherical microphone array data by linear filtering. *EURASIP J. Audio Speech Music Process.*, *2021*(37), 1–11. (*equal contributions). https://doi.org/10.1186/s13636-021-00224-5

Arend, J. M., Amengual Garí, S. V., Schissler, C., Klein, F., & Robinson, P. W. (2021). Six-Degrees-of-Freedom Parametric Spatial Audio Based on One Monaural Room Impulse Response. *J. Audio Eng. Soc.*, *69*(7/8), 557–575. https://doi.org/10.17743/jaes.2021.0009

Pörschmann, C., & **Arend, J. M.** (2021). Investigating phoneme-dependencies of spherical voice directivity patterns. *J. Acoust. Soc. Am.*, *149*(6), 4553–4564. https://doi.org/10.1121/10.0005401

Arend, J. M., Brinkmann, F., & Pörschmann, C. (2021). Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions. *J. Audio Eng. Soc.*, *69*(1/2), 104–117. https://doi.org/10.17743/jaes.2020.0070

Arend, J. M., Liesefeld, H. R., & Pörschmann, C. (2021). On the influence of non-individual binaural cues and the impact of level normalization on auditory distance estimation of nearby sound sources. *Acta Acust.*, *5*(10), 1–21. https://doi.org/10.1051/aacus/2021001

Amengual Garí, S. V., **Arend, J. M.**, Calamia, P., & Robinson, P. W. (2020). Optimizations of the Spatial Decomposition Method for Binaural Reproduction. *J. Audio Eng. Soc.*, *68*(12), 959–976. https://doi.org/10.17743/jaes.2020.0063

Pörschmann, C., Lübeck, T., & **Arend, J. M.** (2020). Impact of face masks on voice radiation. *J. Acoust. Soc. Am.*, *148*(6), 3663–3670. https://doi.org/10.1121/10.0002853

Pörschmann, C., & **Arend, J. M.** (2020). A Method for Spatial Upsampling of Voice Directivity by Directional Equalization. *J. Audio Eng. Soc.*, *68*(9), 649–663. https://doi.org/10.17743/jaes.2020.0033

Lübeck, T., Helmholz, H., **Arend, J. M.**, Pörschmann, C., & Ahrens, J. (2020). Perceptual Evaluation of Mitigation Approaches of Impairments due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data. *J. Audio Eng. Soc.*, *68*(6), 428–440. https://doi.org/10.17743/jaes.2020.0038

Pörschmann*, C., **Arend*, J. M.**, & Brinkmann, F. (2019). Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling. *IEEE/ACM Trans. Audio Speech Lang. Proc.*, *27*(6), 1060–1071. (*equal contributions). https://doi.org/10.1109/TASLP.2019.2908057

Bernschütz, B., Giner, A. V., Pörschmann, C., & **Arend, J. M.** (2014). Binaural Reproduction of Plane Waves With Reduced Modal Order. *Acta Acust. United Ac.*, *100*(5), 972–983. https://doi.org/10.3813/AAA.918777

Peer-Reviewed Conference Publications

Klein, F., Amengual Garí, S. V., **Arend, J. M.**, & Robinson, P. W. (2021). Towards determining thresholds for room divergence: A pilot study on detection thresholds. In *Proc. of the International Conference on Immersive and 3D Audio (I3DA), Bologna, Italy* (pp. 1–7). https://doi.org/10.1109/I3DA48870.2021.9610876

Amengual Garí, S. V., Hassager, H. G., Klein, F., **Arend, J. M.**, & Robinson, P. W. (2021). Towards determining thresholds for the room divergence effect: A pilot study on perceived externalization. In *Proc. of the International Conference on Immersive and 3D Audio (I3DA), Bologna, Italy* (pp. 1–7). https://doi.org/10.1109/I3DA48870.2021.9610835

Bau, D., Lübeck, T., **Arend, J. M.**, Dziwis, D. T., & Pörschmann, C. (2021). Simplifying head-related transfer function measurements: A system for use in regular rooms based on free head movements. In *Proc. of the International Conference on Immersive and 3D Audio (I3DA), Bologna, Italy* (pp. 1–6). https://doi.org/10.1109/I3DA48870.2021.9610890

Dziwis, D. T., Zimmermann, S., Lübeck, T., **Arend, J. M.**, Bau, D., & Pörschmann, C. (2021). Machine Learning-Based Room Classification for Selecting Binaural Reverberation in Augmented Reality Applications. In *Proc. of the International Conference on Immersive and 3D Audio (I3DA), Bologna, Italy* (pp. 1–8). https://doi.org/10.1109/I3DA48870.2021.9610915

Pörschmann, C., Großarth, S., **Arend, J. M.**, Schmitter, S., Schreckenberg, D., & Wunder, K. (2021). Amplitude modulations increase annoyance due to wind turbine noise immission. In *Proc. of Inter-Noise, Washington, DC, USA* (pp. 4048-4057). https://doi.org/10.3397/IN-2021-2589

Dziwis, D. T., **Arend, J. M.**, Lübeck, T., & Pörschmann, C. (2021). IVES – Interactive Virtual Environment System: A Modular Toolkit for 3D Audiovisual Composition in Max. In *Proc. of the 18th Sound and Music Computing Conference (SMC), Torino, Italy* (pp. 330–337).

Lübeck, T., Helmholz, H., **Arend, J. M.**, Pörschmann, C., & Ahrens, J. (2020). Perceptual Evaluation of Mitigation Approaches of Impairments due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data: Dry Acoustic Environments. In *Proc. of the 23rd International Conference on Digital Audio Effects (DAFx2020), Vienna, Austria* (pp. 250–257).

Lübeck, T., Pörschmann, C., & **Arend, J. M.** (2020). Perception of direct sound, early reflections, and reverberation in auralizations of sparsely measured binaural room impulse responses. In *Proc. of the AES International Conference on Audio for Virtual and Augmented Reality (AVAR), Redmond, WA, USA* (pp. 1–10).

Pörschmann, C., **Arend, J. M.**, Bau, D., & Lübeck, T. (2020). Comparison of Spherical Harmonics and Nearest-Neighbor based Interpolation of Head-Related Transfer Functions. In *Proc. of the AES International Conference on Audio for Virtual and Augmented Reality (AVAR), Redmond, WA, USA* (pp. 1–10).

Pörschmann, C., & **Arend, J. M.** (2019). How positioning inaccuracies influence the spatial upsampling of sparse head-related transfer function sets. In *Proc. of the International Conference on Spatial Audio (ICSA), Ilmenau, Germany* (pp. 47–54). https://doi.org/10.22032/dbt.39952

Pörschmann, C., **Arend, J. M.**, & Gillioz, R. (2019). How wearing headgear affects measured head-related transfer functions. In *Proc. of the EAA Spatial Audio Signal Processing Symposium, Paris, France* (pp. 49–54). https://doi.org/10.25836/sasp.2019.27

Arend, J. M., & Pörschmann, C. (2019). Spatial upsampling of sparse head-related transfer function sets by directional equalization – Influence of the spherical sampling scheme. In *Proc. of the 23rd International Congress on Acoustics (ICA), Aachen, Germany* (pp. 2643–2650). https://doi.org/10.18154/RWTH-CONV-238939

Pörschmann, C., **Arend, J. M.**, & Brinkmann, F. (2019). Spatial upsampling of individual sparse head-related transfer function sets by directional equalization. In *Proc. of the 23rd International Congress on Acoustics (ICA), Aachen, Germany* (pp. 4870–4877). https://doi.org/10.18154/RWTH-CONV-239484

Arend, J. M., Lübeck, T., & Pörschmann, C. (2019). A Reactive Virtual Acoustic Environment for Interactive Immersive Audio. In *Proc. of the AES International Conference on Immersive and Interactive Audio (IIA), York, UK* (pp. 1–10).

Pörschmann, C., & **Arend, J. M.** (2019). Obtaining Dense HRTF Sets from Sparse Measurements in Reverberant Environments. In *Proc. of the AES International Conference on Immersive and Interactive Audio (IIA), York, UK* (pp. 1–10).

Pörschmann, C., **Arend, J. M.**, & Stade, P. (2018). Investigations on the Impact of Distance Cues in Virtual Acoustic Environments. In *Proc. of the 30th Tonmeistertagung – VDT International Convention, Cologne, Germany* (pp. 229–236).

Pörschmann, C., Stade, P., & **Arend, J. M.** (2017). Binauralization of Omnidirectional Room Impulse Responses – Algorithm and Technical Evaluation. In *Proc. of the 20th International Conference on Digital Audio Effects (DAFx17), Edinburgh, UK* (pp. 345–352).

Stade, P., **Arend, J. M.**, & Pörschmann, C. (2017). Perceptual Evaluation of Synthetic Early Binaural Room Impulse Responses Based on a Parametric Model. In *Proc. of the 142nd AES Convention, Berlin, Germany* (pp. 1–10).

Arend, J. M., & Pörschmann, C. (2016). Audio Watermarking of Binaural Room Impulse Responses. In *Proc. of the AES International Conference on Headphone Technology, Aalborg, Denmark* (pp. 1–8).

Stade, P., & **Arend, J. M.** (2016). Perceptual Evaluation of Synthetic Late Binaural Reverberation Based on a Parametric Model. In *Proc. of the AES International Conference on Headphone Technology, Aalborg, Denmark* (pp. 1–8).

Conference Publications

Lübeck, T., **Arend, J. M.**, & Pörschmann, C. (2021). A High-Resolution Spatial Room Impulse Response Database. In *Proc. of the 47th DAGA, Vienna, Austria* (pp. 1604-1607).

Amengual Garí, S. V., **Arend, J. M.**, Calamia, P., & Robinson, P. W. (2020). Optimizing the Spatial Decomposition Method for Binaural Rendering. In *Proc. of Forum Acusticum, Lyon, France* (pp. 1213–1220). https://doi.org/10.48465/fa.2020.0688

Lübeck, T., **Arend, J. M.**, Helmholz, H., Ahrens, J., & Pörschmann, C. (2020). Comparison of Mitigation Approaches of Spatial Undersampling Artifacts in Spherical Microphone Array Data Auralizations. In *Proc. of the 46th DAGA, Hannover, Germany* (pp. 1–4).

Pörschmann, C., & **Arend, J. M.** (2020). Analyzing the Directivity Patterns of Human Speakers. In *Proc. of the 46th DAGA, Hannover, Germany* (pp. 1141–1144).

Arend, J. M., & Pörschmann, C. (2019). Synthesis of Near-Field HRTFs by Directional Equalization of Far-Field Datasets. In *Proc. of the 45th DAGA, Rostock, Germany* (pp. 1454–1457).

Pörschmann, C., & **Arend, J. M.** (2019). A Method for Spatial Upsampling of Directivity Patterns of Human Speakers by Directional Equalization. In *Proc. of the 45th DAGA, Rostock, Germany* (pp. 1458–1461).

Lübeck, T., **Arend, J. M.**, & Pörschmann, C. (2019). HMD-based Virtual Environments for Localization Experiments. In *Proc. of the 45th DAGA, Rostock, Germany* (pp. 1116–1119).

Dziwis, D. T., Lübeck, T., **Arend, J. M.**, & Pörschmann, C. (2019). Development of a 7th Order Spherical Microphone Array for Spatial Audio Recording. In *Proc. of the 45th DAGA, Rostock, Germany* (pp. 883–885).

Lübeck, T., **Arend, J. M.**, & Pörschmann, C. (2018). A Real-Time Application for Sound Source Localization Inside a Spherical Microphone Array. In *Proc. of the 44th DAGA, Munich, Germany* (pp. 319–322).

Arend, J. M., Stade, P., & Pörschmann, C. (2017). Binaural reproduction of self-generated sound in virtual acoustic environments. *Proc. of Meetings on Acoustics*, *30*(015007), 1-14. https://doi.org/10.1121/2.0000574

Stade, P., **Arend, J. M.**, & Pörschmann, C. (2017). A parametric model for the synthesis of binaural room impulse responses. *Proc. of Meetings on Acoustics*, *30*(015006), 1-12. https://doi.org/10.1121/2.0000573

Pörschmann, C., Stade, P., & **Arend, J. M.** (2017). Binaural auralization of proposed room modifications based on measured omnidirectional room impulse responses. *Proc. of Meetings on Acoustics*, *30*(015012), 1-12. https://doi.org/10.1121/2.0000622

Pörschmann, C., **Arend, J. M.**, & Neidhardt, A. (2017). A Spherical Near-Field HRTF Set for Auralization and Psychoacoustic Research. In *Proc. of the 142nd AES Convention, Berlin, Germany* (pp. 1–5).

Arend, J. M., Stade, P., & Pörschmann, C. (2017). A System for Binaural Reproduction of Self-Generated Sound in VAEs. In *Proc. of the 43rd DAGA, Kiel, Germany* (pp. 271–274).

Pörschmann, C., **Arend, J. M.**, & Stade, P. (2017). Influence of head tracking on distance estimation of nearby sound sources. In *Proc. of the 43rd DAGA, Kiel, Germany* (pp. 1065–1068).

Stade, P., & **Arend, J. M.** (2017). Synthetic Reflections for Binaural Rendering using Sound Field Analysis. In *Proc. of the 43rd DAGA, Kiel, Germany* (pp. 1089–1092).

Arend, J. M., Neidhardt, A., & Pörschmann, C. (2016). Measurement and Perceptual Evaluation of a Spherical Near-Field HRTF Set. In *Proc. of the 29th Tonmeistertagung – VDT International Convention, Cologne, Germany* (pp. 356–363).

Stade, P., & **Arend, J. M.** (2016). A Perception-Based Parametric Model for Synthetic Late Binaural Reverberation. In *Proc. of the 42nd DAGA, Aachen, Germany* (pp. 63–66).

Ebelt, M. D., **Arend, J. M.**, & Pörschmann, C. (2016). Influences of the Floor Reflection on Auditory Distance Perception. In *Proc. of the 42nd DAGA, Aachen, Germany* (pp. 1467–1469).

Stade, P., **Arend, J. M.**, & Goebels, K. (2014). Anwendung neuer Methoden zur raumakustischen Analyse in einem Regieraum der WDR-Hörspielstudios. In *Proc. of the 40th DAGA, Oldenburg, Germany* (pp. 602–603).

Various Publications

Pörschmann, C., **Arend, J. M.**, & Wunder, K. (2022). Einfluss von Immissionspegel und Amplitudenmodulation auf die Lästigkeit von Windenergieanlagen/Influence of immission level and amplitude modulation on the annoyance of wind turbines. *Lärmbekämpfung*, *17*(01), 17–21. https://doi.org/10.37544/1863-4672-2022-01-19

Schmitter, S., Di Loro, A. A., Pörschmann, C., **Arend, J. M.**, Großarth, S., & Schreckenberg, D. (2021). Geräuschwirkungen bei der Nutzung von Windenergie an Land. *Akustik Journal*, *21*(03), 16–30.

Arend, J. M., Neidhardt, A., & Pörschmann, C. (2017). Measurement and Evaluation of a Near-Field HRTF Set. *VDT Magazin*, *2017*(1), 52–55.

Open Software and Data Contributions

Besides the scientific research papers included in the list of publications, the author contributed the following open software and data to the scientific community, which are also considered part of this thesis.

Lübeck, T., **Arend, J. M.**, & Pörschmann, C. (2021). A High-Resolution Spatial Room Impulse Response Database. https://doi.org/10.5281/zenodo.5031335

Arend*, J. M., Lübeck*, T., & Pörschmann*, C. (2021). SMATBIN – A Matlab toolbox to calculate linear filters for efficient binaural rendering of spherical microphone array data. (*equal contributions). https://github.com/AudioGroupCologne/SMATBIN

Arend, J. M., Amengual Garí, S. V., Schissler, C., Klein, F., & Robinson, P. W. (2021). Paraspax – A Matlab toolbox for parametric spatial audio with six degrees of freedom. https://github.com/facebookresearch/Paraspax

Pörschmann, C., & **Arend, J. M.** (2021). Full-spherical voice directivity measurements for 13 subjects and 8 phonemes. https://doi.org/10.5281/zenodo.4739183

Pörschmann, C., Lübeck, T., & **Arend, J. M.** (2020). Full-spherical directivity measurements of the HEAD acoustics HMS II.3 head and mouth simulator with different face masks. https://doi.org/10.5281/zenodo.3952320

Pörschmann, C., **Arend, J. M.**, & Gillioz, R. (2019). Spherical Headgear HRIR Compilation of the Neumann KU100 and the HEAD acoustics HMS II.3. https://doi.org/10.5281/zenodo.3928465

Arend, J. M., Pörschmann, C., & Brinkmann, F. (2019). SUpDEq – A Matlab toolbox for spatial upsampling and processing of HRTFs. https://github.com/AudioGroupCologne/SUpDEq

Arend, J. M., Neidhardt, A., & Pörschmann, C. (2016). Near-field HRTF database of the Neumann KU100 dummy head. https://doi.org/10.5281/zenodo.3928399