Variable-Orientation Auralization based on Room Response Measurements Involving Directivity

Binaural and Higher-Order Ambisonic Methods

DISSERTATION

submitted by

Markus Zaunschirm

(matriculation number: 0530581)

submitted to

University of Music and Performing Arts Graz
PhD Program (Sound and Music Computing, V 094 750)

Supervisors

Prof. Dr. Robert Höldrich
Prof. Dr. Gerhard Eckel

External Reviewer

Prof. Dr.-Ing. Sascha Spors

Graz, July 27, 2020

Abstract

An interactive and flexible measurement-based auralization of an acoustic scenery benefits from a separation into source-, room-, and receiver-dependent modules. This thesis presents a room description that facilitates such a modularity: the source-and-receiver-directional Ambisonics room impulse response (SRD ARIR) capture and processing approach. In its most hardware-efficient implementation, the SRD ARIR relies on a small set of RIRs measured between a first-order source and a first-order receiver. In order to facilitate the auralization of sources with higher-order directivity, the Ambisonic spatial decomposition method (ASDM) is employed to enhance the directional resolution, i.e. to upscale the first-order resolution of the measurements to higher orders. In the Ambisonics domain, the SRD ARIR interfaces seamlessly with the source and receiver directivities, which are typically available in Ambisonics as well.

On the receiver side, this thesis presents perpetually motivated modifications of the head-related transfer functions (HRTFs) that radically improve binaural rendering of Ambisonic signals. The methods either employ a frequency-dependent HRTF time alignment in pre-processing or use a magnitude-least-squares optimization where a phase-match at high frequencies is disregarded in favor of a magnitude match. Both renderers optionally include an interaural covariance correction that enforces optimal rendering of diffuse fields with only small impact when rendering particular free fields. Results from the presented listening experiments indicate that already an order of three allows for high-quality rendering.

Measurement-based auralization does not exclusively rely on Ambisonics. Especially if modularity is not required, auralization based on multiple-orientation binaural room impulse responses (MOBRIRs) is a popular alternative. This thesis discusses the optimal MOBRIR resolution that allows for high-quality variable-orientation rendering while keeping the measurement effort low. The results from listening experiments comparing various orientation resolutions indicate that the optimum is found for a resolution of 15° or finer.

The proposed SRD ARIR method is perceptually evaluated in listening experiments where a MOBRIR-based auralization is employed as a reference condition. For both the MOBRIR- and the SRD ARIR-based auralization, the icosahedral loudspeaker array (IKO) was employed as directional source of well-studied perceptual effects. The results of the listening experiments indicate results of similar quality when comparing the proposed SRD ARIR method to alternative rendering methods, when using measurements taken in the same acoustic environment.

Kurzfassung

Eine interaktive, flexible und auf Messdaten basierende Auralisation einer akustischen Szene profitiert von einer Trennung in die Quell-, Raum- und Empfängermodule. In dieser Arbeit wird eine Mess- und Verarbeitungsstrategie des Raummoduls vorgestellt, die eine solche Modularität ermöglicht: die Source-and-Receiver-Directional Ambisonics Raumimpulsantwort (SRD ARIR). In ihrer effizientesten (Hardware) Implementierung stützt sich die SRD-ARIR auf einen begrenzten Satz von RIRs, die zwischen einer Quelle und einem Empfänger von jeweils erster Ordnung gemessen werden. Um die Auralisation von Quellen mit Richtwirkungen höherer Ordnung zu ermöglichen, wird die Ambisonic Spatial Decomposition Method (ASDM) verwendet, um die Richtungsauflösung zu erhöhen, d.h. die Auflösung der Messungen erster Ordnung auf höhere Ordnungen hoch zu skalieren. In der Ambisonics-Domäne lässt sich die SRD-ARIR nahtlos mit der Quell- und Empfänger-Richtcharakteristik verbinden, die typischerweise auch in Ambisonics zur Verfügung stehen.

Auf der Empfängerseite werden in dieser Arbeit Ansätze zur binauralen Wiedergabe von Ambisonics Signalen vorgestellt, die diese radikal verbessert. Es wird ein frequenzabhängiger Laufzeitabgleich in der Vorverarbeitung der Außenohrübertragungsfunktion (HRTF) oder eine Optimierung der kleinsten Fehlerquadrate, die eine Phasenanpassung zugunsten einer Amplitudenanpassung bei hohen Frequenzen vernachlässigt, vorgeschlagen. Beide Renderer enthalten optional eine interaurale Kovarianzkorrektur, die eine optimale Wiedergabe von Diffusfeldern bei nur geringer Auswirkung auf die Wiedergabe vom Direktschall erzwingt. Die Ergebnisse der durchgeführten Hörversuche deuten darauf hin, dass die neuen Ansätze ab einer Ambisonics Ordnung von drei eine quasi-transparente und qualitativ hochwertige Wiedergabe ermöglichen.

Auf Messdaten basierende Auralisation setzt jedoch nicht ausschließlich auf Ambisonics. Insbesondere dann, wenn Modularität nicht erforderlich ist, ist die Auralisation auf der Basis von multi-orientation-binaural Raumimpulsantworten (MOBRIRs) eine gängige Alternative. In dieser Arbeit wird die optimale MOBRIR-Auflösung diskutiert, die eine qualitativ hochwertige dynamische Binauralsynthese ermöglicht und gleichzeitig den Messaufwand gering hält. Die Ergebnisse aus Hörversuchen zum Vergleich verschiedener Orientierungsauflösungen zeigen, dass das Optimum bei einer Auflösung von 15° liegt.

Um die vorgeschlagene SRD-ARIR-Methode wahrnehmungstechnisch zu evaluieren werden Hörversuche mit MOBRIR-basierter Auralisation als Referenz durchgeführt. Sowohl für die MOBRIR- als auch für die SRD-ARIR-basierte Auralisation wurde das Ikosaeder-Lautsprecherarray (IKO) als Quelle höherer Ordnung eingesetzt. Die Ergebnisse des Hörversuchs zeigen, dass die vorgeschlagene SRD-ARIR-Methode vergleichbare Qualität wie die Alternativen, die auf Messungen in derselben akustischen Umgebung basieren, erreicht.

Contents

1 Introduction

Auralization, as defined in [P9,ref 31][1], is the technique of creating audible sound files from numerical (simulated, measured, or synthesized) data. Typically, measurement-based auralization employs binaural room impulse responses (BRIRs), which are measured between an omnidirectional source and a dummy head. While this approach is known to yield high audio quality and convincing realism [P8,ref 1,ref 4], it has a draw back: both, the source, and receiver directivity are fixed during measurements and can't be exchanged during post-production or interactive playback. Thus, auralization based on BRIR measurements does neither allow for listener personalisation, i.e. individualization to a listener's head related impulse responses (HRIRs), nor for insertion of an arbitrary source directivity (e.g. of a musical instrument).

In order to facilitate exchangeable or variable source and receiver directivities for a modular and interactive measurement-based auralization, a separation into the source, room, and receiver module is beneficial. Typically, source and receiver directivities are measured with a surrounding microphone or loudspeaker array, respectively [e.g. P5]. For their continuous directional interpolation, a representation of those directivities in the Ambisonics domain is quite common. Both require a generalized representation of the room that interfaces with those directivity measurements in Ambisonics.

A concept and measurement method for such a room description with Ambisonic directivity interfaces is presented in chapter 2, and the response format source-and-receiver directional room impulse response in Ambisonics (SRD ARIR) is introduced [P9]. For efficient measurements of the SRD ARIR, first-order arrays are employed at both the source- and receiver-side, respectively. Those first-order measurements are then processed and upscaled using the spatial decomposition method in the Ambisonics domain (ASDM) [P9,ref 22-ref 23] at both sides. Throughout the chapter and [P9], the question: *Can auralization of a highly directional source in a room succeed if it employs a room impulse response (RIR) measurement or simulation relying on a first-order directional source, only?*, is answered. The results from the presented listening experiment, where the proposed upscaled SRD ARIR is also compared against a directly measured higher-order SRD ARIR as described in [P1], encourage the conclusion.

The findings and studies discussed in the other chapters of this thesis present a more detailed description of individual processing steps, support a deeper technical understanding and offer in-depth perceptual studies. This thesis is structured as follows.

Chapter 3 presents an in-depth overview of the icosahedral loudspeaker array (IKO[2]), a compact spherical loudspeaker array consisting of 20 loudspeakers placed on the surfaces of an icosahedron. The IKO's technical background, staging, virtualization, and its beamforming and control are presented in section 3.1 [P2]. The proposed beamforming

[1] refers to reference 31 in publication P9.
[2] `https://iko.sonible.com/en.html`

approach limits the excursion of the IKO's loudspeakers and is based on Laser Doppler vibrometry measurements. While parts of the underlying measurements and resulting beam patterns (evaluated using pressure measurements of a surrounding microphone array) are already discussed in section 3.1, a more complete measurement and evaluation repository including also the data of the more recent IKO versions is presented section 3.2 [P5]. Due to its well-studied perceptual effects [P9,ref 5,ref 7], its well-defined Ambisonics beamforming [P2], and its technical documentation (high resolution directivity measurements [P5]), the IKO is employed as the source with controllable directivity throughout the listening experiment presented in chapter 2.

In a headphone based auralization scenario, the receiver, i.e. a listener with two ears, is described by its individual HRIRs which implicitly contain the cues accessible to the human auditory system to perceive sound from a certain direction and distance, with a certain source width, envelopment, or spaciousness [P8,ref 2,ref 3]. When employing the SRD ARIR, the playback can be personalized to the listener by using an individualized Ambisonics to binaural renderer. Chapter 4 discusses optimization and pre-processing strategies yielding such efficient binaural renderers, i.e. a high-quality but low-order representation of HRIRs in Ambisonics. Section 4.1 [P3] discusses a binaural renderer that is computed using a frequency-dependent time alignment of HRIRs followed by a minimization of the squared error subject to a diffuse-field covariance matrix constraint. Section 4.2 [P4] describes and presents the MagLS (magnitude-least-squares) renderer as a further development that managed to get widely adopted in software applications[3]. Its filters are designed using a magnitude-least-squares optimization that disregards a phase match in favor of an improved HRTF magnitude at high frequencies. Both renderers include an interaural covariance correction that offers to render diffuse fields consistently.

Chapter 5 discusses interactive head-tracked (dynamic) binaural rendering based on multi-orientation BRIRs (MOBRIRs). The acquisition of a high-resolution MOBRIR set can be tedious, and individualized auralization requires separate measurements with each listener in each room. Thus, efficient alternatives propose BRIR individualization by measuring the listener-dependent (HRIRs) and room-dependent (RIRs) parts separately [P8,ref 12-ref 14]. Section 5.1 [P6] presents a study dealing with the BRIR synthesis from first-order measurements. This section focuses on an evaluation based on technical measures and presents a perceptual study employing static binaural rendering. A more in-depth perceptual study employing also dynamic rendering, i.e. adapt to the head movements of a listener, is discussed in section 5.3[P8]. Section 5.2 [P7] deals with MOBRIR interpolation and discusses the required angular resolution for high quality dynamic rendering. The underlying implementation and other parameter settings, e.g. processing block size, and cross-over frequency, are discussed in this section as well. Note that the perceptual studies presented throughout the chapters 2, and 5 involve interactive dynamic binaural rendering and a reference condition which is based on the findings presented in section 5.2.

Finally, the concluding remarks are presented in chapter 6.

[3] `https://plugins.iem.at/`, `http://research.spa.aalto.fi/projects/sparta_vsts/plugins.html`

List of Publications

This thesis consists of an introduction and the publications which are listed in a chronological order here. Please note that the author's contributions to each publication are given at the beginning of the corresponding sections.

[P1] M. Zaunschirm, M. Frank, and F. Zotter. An interactive virtual icosahedral loudspeaker array. *Proceedings of the DAGA*, Aachen, 2016.

[P2] F. Zotter, M. Zaunschirm, M. Frank, and M. Kronlachner. A Beamformer to Play with Wall Reflections: The Icosahedral Loudspeaker. *Computer Music Journal*, 41 (3), 2017. ISSN 15315169. doi:10.1162/comj_a_00429.

[P3] M. Zaunschirm, C. Schörkhuber, and R. Höldrich. Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint. *J. Acoust. Soc. Am., 143(6):3616–3627.* 2018. doi:10.1121/1.5040489.

[P4] C. Schörkhuber, M. Zaunschirm, and R. Höldrich. Binaural Rendering of Ambisonic Signals via Magnitude Least Squares. *Proceedings of the DAGA*, 44:339–342. 2018.

[P5] F. Schultz, M. Zaunschirm, and F. Zotter.. Directivity and electro-acoustic measurements of the IKO. *Audio Engineering Society Convention e-Brief 144*, Milano. 2018.

[P6] M. Zaunschirm, M. Frank, and F. Zotter. BRIR synthesis using first-order microphone arrays. *Convention of the Audio Eng. Soc. 144*, Milano, pages 1–10. 2018.

[P7] M. Zaunschirm, M. Frank, and F. Zotter. Perceptual Evaluation of Variable-Orientation Binaural Room Impulse Response Rendering. *Conference of the Audio Eng. Soc.: 2019 AES International Conference on Immersive and Interactive Audio*, York, UK.

[P8] M. Zaunschirm, M. Frank, and F. Zotter. Binaural Rendering with Measured Room Responses: First-Order Ambisonic Microphone vs. Dummy Head. *Applied Sciences*, 10(5). 2020.

[P9] M. Zaunschirm, F. Zagala, and F. Zotter. Auralization of High-Order Directional Sources from First-Order RIR Measurements. *Applied Sciences*, 10(11):3747. 2020.

2

Auralization of High-Order Directional Sources

2.1 Auralization of High-Order Directional Sources from First-Order RIR Measurements

This work was published as:

M. Zaunschirm, F. Zagala, and F. Zotter. (2020). Auralization of High-Order Directional Sources from First-Order RIR Measurements. *Applied Sciences*, 10(11):3747.

The idea and concept of this article were outlined by me, the first author, with help from the third author. I wrote the original draft of the manuscript with periodical contributions from the third and second author. The revision and editing was done by me with help from the third and second author. I did most of the programming, and prepared the samples for the listening experiment. I programmed and designed the listening experiment in Unity with periodic contributions from the third author. The listening experiment was conducted and evaluated by the second author, with periodic contributions from the third author and me.

applied sciences

MDPI

Article

Auralization of High-Order Directional Sources from First-Order RIR Measurements

Markus Zaunschirm [1,*], Franck Zagala [1,2,3] and Franz Zotter [1]

[1] Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Inffeldgasse 10/III, 8010 Graz, Austria; franck.zagala@ircam.fr; zotter@iem.at
[2] Institut Jean Le Rond, CNRS, UMR 7190, Sorbonne Université ∂'Alembert, 75005 Paris, France
[3] Sciences et Technologies de la Musique et du Son (STMS) - Sorbonne Université, IRCAM, CNRS, 75004 Paris, France
* Correspondence: zaunschirm@iem.at or markus.zaunschirm@atmoky.com

Received: 17 April 2020; Accepted: 26 May 2020; Published: 28 May 2020

Abstract: Can auralization of a highly directional source in a room succeed if it employs a room impulse response (RIR) measurement or simulation relying on a first-order directional source, only? This contribution presents model and evaluation of a source-and-receiver-directional Ambisonics RIR capture and processing approach (SRD ARIR) based on a small set of responses from a first-order source to a first-order receiver. To enhance the directional resolution, we extend the Ambisonic spatial decomposition method (ASDM) to upscale the first-order resolution of both source and receiver to higher orders. To evaluate the method, a listening experiment was conducted based on first-order SRD-ARIR measurements, into which the higher-order directivity of icosahedral loudspeaker's (IKO) was inserted as directional source of well-studied perceptual effects. The results show how the proposed method performs and compares to alternative rendering methods based on measurements taken in the same acoustic environment, e.g., multiple-orientation binaural room impulse responses (MOBRIRs) from the physical IKO to the KU-100 dummy head, or higher-order SRD ARIRs from IKO to em32 Eigenmike. For optimal externalization, our experiments exploit the benefits of virtual reality, using a highly realistic visualization on head-mounted-display, and a user interface to report localization by placing interactive visual objects in the virtual space.

Keywords: measurement-based auralization; room impulse response measurement; source directivity; ASDM upscaling; dynamic binaural rendering; BRIR measurements; audio for augmented reality; psychoacoustics

1. Introduction

A modular and interactive measurement-based auralization of an acoustic environment benefits from a separation into its source-dependent, room-dependent, and receiver-dependent parts. Typically, the room-dependent part is characterized by a point-to-point room impulse response (RIR), which often assumes that source and receiver are both omnidirectional [1]. However, employing variable source and receiver directivities during auralization requires a more flexible room description that facilitates interfacing between the three parts.

Why source directivity matters : Otondo and Rindel [2] demonstrated that room acoustics parameters change with source directivity, and results from listening experiments indicate that the resulting loudness, reverberance, and clarity changes induced by directivity are perceived by listeners. Vigeant et al. [3] found that including source directivity can increase the realism of auralization results. Latinen et al. [4] showed that the source directivity can be used to alter the direct-to-reverberant ratio, which strongly correlates with perceived distance of a source. Another study employing a source

of controllable higher-order directivity by Wendt et al. [5] showed that both the auditory source distance and the apparent source width are influenced by the directivity. Ronsse and Wang [6] found source directivity to modify clarity, localization, and timbre. In terms of localization, Wendt et al. [7] investigated how beam-formed source directivity of an icosahedral loudspeaker (IKO) produces auditory events that can be shifted between physical source and wall reflections, or follow traceable trajectories. Wang and Vigeant [8] demonstrated the influence of source directivity on reverberation time and clarity, and they found a clear effect of source directivity in their auralization experiment.

Why receiver directivity matters: Higher-order receiver directivities have been recently proved useful to characterize room acoustical measurements, see, e.g., in [9,10], and moreover, e.g., to identify the sound field isotropy in various reverberant rooms [11,12]. In the targeted auralization scenario, the receiver is obviously a listener and thus the various sound propagation paths arriving at the ears are weighted by their head related transfer functions (HRTFs). In order to employ an arbitrary receiver directivities during auralization or postprocessing, e.g., individualized HRTFs, the measured room-dependent part also has to be generic and has to facilitate higher-order receiver directivities.

Why we suggest Ambisonics: Directivities are typically measured under anechoic conditions, with a microphone array surrounding the source, or a loudspeaker array surrounding the receiver, respectively. For comparability and a unified directional interpolation, a representation of those directivities in terms of spherical harmonic expansion coefficients is beneficial, see, e.g., in [13,14]. Consequently, a generalized representation of the room-dependent part that interfaces with both the source and receiver directivities should also be expanded in spherical harmonics, i.e. represented in Ambisonics. Furthermore, Oberem et al. [15] found that dynamic binaural rendering (incorporating head rotations) significantly improves localization accuracy. Moreover, using Ambisonics has its benefits in facilitating dynamic rendering efficiently, as it implements dynamic sound scene rotation by a time-variant matrix multiplication [16,17], whereas the convolution with MagLs [18] HRIRs in the spherical harmonics domain remains time-invariant.

SRD ARIR: According to the above considerations, we propose a source-and-receiver-directional (SRD) higher-order Ambisonic room impulse response (ARIR) as representation of the room-dependent parts of auralization. Measuring such a SRD ARIR requires high-order spherical microphone and loudspeaker arrays, which are recently used for room acoustical measurements or propagation path identification [19,20] or as well for studying the concert hall preference [14].

Alternatively to measuring with high-order arrays, we introduced measurements with greatly reduced hardware effort in [21], where first-order source and receiver arrays are employed. The desired higher-order resolution at the receiver is obtained through our Ambisonic spatial decomposition method (ASDM) [22] that is based on Tervo et al. [23]. As an extension involving the source side, we propose the SRD ARIR algorithm that assigns a highly resolved direction of departure (DOD) and direction of arrival (DOA) to each sample of the omnidirectional-source-to-omnidirectional-receiver RIR.

Contents: In this contribution we present the concept and processing steps of a measurement-based auralization for high-order directional sources from hardware-efficient first-order measurements in Section 2. To facilitate an interactive real-time auralization, the source- and receiver-directivities, as well as the SRD ARIR are represented in the Ambisonic domain. The design and implementation of a comparative listening experiment are discussed in Section 3.2. For the sake of reproducibility and generalizability we employed the 20-channel icosahedral loudspeaker array (IKO, https://iko.sonible.com/) as a source with well-described controllable directivity [24]. The five measurement-based auralization techniques under test are described in Section 3.1 and include (i) dummy head BRIR rendering as defined in [22], (ii) rendering with MIMO RIRs as defined in [25], (iii) rendering using multi ASDM RIRs (upscaled for each individual IKO transducer), (iv) rendering using the upscaled SRD ARIR and a generic 3rd order directivity, and (v) rendering using the upscaled SRD ARIR with the real IKO directivity. The underlying measurements were taken in the György Ligeti Saal (V = 5630 m^3, $T_{60} = 1.4$ s.). A statistical analysis and discussion of the ratings is given in Section 3.3.

Appl. Sci. **2020**, *10*, 3747

2. Auralization of Arbitrary Source Directivity from Measurements

The block diagram of an auralization scenario employing the SRD ARIR is shown in Figure 1 and is similar to that in [26]. Source and receiver directivities are interfaced with the room through Ambisonic input and output signals. Here, the receiver directivity is represented by HRTFs, and a state-of-the-art binaural renderer, e.g., the MagLS renderer as outlined in [18], is used for obtaining the signals that are fed to headphones.

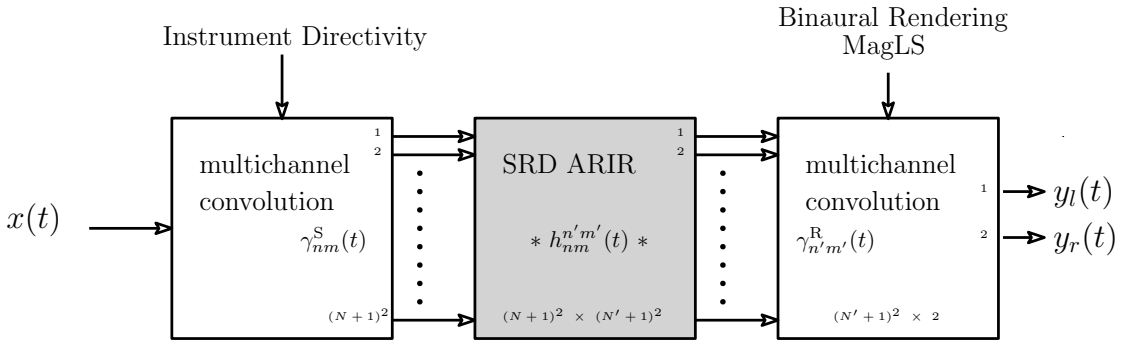


Figure 1. Auralization scenario using the source and receiver directional (SRD) ARIR (in the Ambisonics domain). Here, $x(t)$ is the source signal, $h^{n'm'}_{nm}(t)$ is the SRD ARIR, and $\gamma^{\mathrm{S}}_{nm}(t)$ and $\gamma^{\mathrm{R}}_{n'm'}(t)$ are the directional impulse responses of the source and receiver, e.g., ear directivity, respectively.

The concept of the SRD ARIR as well as its use for auralization is described in Section 2.1. The proposed hardware efficient (low order) SRD ARIR measurement method and the upscaling to higher orders is discussed in Section 2.2.

2.1. Theory behind Source and Receiver Directional (SRD) RIRs

Based on the image source method [27], or more generally, the geometrical theory of diffraction [28], physically consistent room acoustic models including edge diffraction can be devised based on geometric sound propagation paths, see, e.g., in [29–31]. Accordingly, we may write any source-and-receiver-directional room impulse response (SRD RIR) as the sum of discrete propagation paths of the index i

$$h\left(\boldsymbol{\theta}_{\mathrm{R}}, t, \boldsymbol{\theta}_{\mathrm{S}}\right) = \sum_i \frac{\delta(\boldsymbol{\theta}_{\mathrm{R}} - \boldsymbol{\theta}_{\mathrm{R},i})\, \mathrm{a}_i\, \delta(t-\tau_i)\delta(\boldsymbol{\theta}_{\mathrm{S}} - \boldsymbol{\theta}_{\mathrm{S},i})}{\mathrm{r}_i}, \tag{1}$$

where each path is characterized by a direction of arrival (DOA) and departure (DOD) denoted as $\boldsymbol{\theta}_{\mathrm{R},i}$ and $\boldsymbol{\theta}_{\mathrm{S},i}$, respectively; an arrival time $\tau_i = \frac{\mathrm{r}_i}{c}$; its geometric length r_i; and its attenuation a_i through reflection and diffraction on rigid or sound-soft surfaces. For complex surface impedances, multiplication by a_i theoretically becomes convolution by an impulse response $\mathrm{a}_i \rightarrow \mathrm{a}_i(t)*$ or it can be expanded to additional paths in discrete-time processing, as preferred here. All vectors describing continuous or discrete directions $\boldsymbol{\theta}$ are denoted as unit direction vectors $\boldsymbol{\theta} = [\cos(\varphi)\sin(\vartheta), \sin(\varphi)\sin(\vartheta), \cos(\vartheta)]^{\mathrm{T}}$, with φ denoting the azimuth and ϑ the zenith angle; labels S and R refer to source or receiver.

We assume a signal $x(t)$ that gets emitted by a source with the directivity $g_{\mathrm{S}}(\boldsymbol{\theta}_{\mathrm{S}})$ and gets picked up with the receiver directivity $g_{\mathrm{R}}(\boldsymbol{\theta}_{\mathrm{R}})$. The resulting signal $y(t)$ is described by the convolution with the following impulse response,

$$y(t) = x(t) * h(t) \qquad \text{with } h(t) = \int_{\mathbb{S}^2}\int_{\mathbb{S}^2} g_{\mathrm{R}}(\boldsymbol{\theta}_{\mathrm{R}})\, h\left(\boldsymbol{\theta}_{\mathrm{R}}, t, \boldsymbol{\theta}_{\mathrm{S}}\right) g_{\mathrm{S}}(\boldsymbol{\theta}_{\mathrm{S}})\, \mathrm{d}\boldsymbol{\theta}_{\mathrm{R}}\, \mathrm{d}\boldsymbol{\theta}_{\mathrm{S}}. \tag{2}$$

The RIR $h(t)$ is obtained by weighting the SRD RIR $h\left(\boldsymbol{\theta}_{\mathrm{R}}, t, \boldsymbol{\theta}_{\mathrm{S}}\right)$ with both the source and receiver directivities $g_{\mathrm{S}}(\boldsymbol{\theta}_{\mathrm{S}})$ and $g_{\mathrm{R}}(\boldsymbol{\theta}_{\mathrm{R}})$, assuming they are frequency-independent; for frequency-dependent

directivities, multiplication by the directivities is replaced by convolutions with the directional impulse responses of source $g_\mathrm{S}(\boldsymbol{\theta}_\mathrm{S}) \rightarrow *g_\mathrm{S}(t, \boldsymbol{\theta}_\mathrm{S})$ and receiver $g_\mathrm{R}(\boldsymbol{\theta}_\mathrm{R}) \rightarrow g_\mathrm{R}(t, \boldsymbol{\theta}_\mathrm{R})*$, respectively.

In the Ambisonic domain: Equation (1) is transformed into the spherical harmonic domain by integrating either dependency on the variable sending and receiving direction over the spherical harmonics. As a result, the spherical delta functions are replaced by spherical harmonics (SH) evaluated at either DOA ($\boldsymbol{\theta}_{\mathrm{R},i}$) or DOD ($\boldsymbol{\theta}_{\mathrm{S},i}$) of the respective propagation path i. We get

$$h\left(\boldsymbol{\theta}_\mathrm{R}, t, \boldsymbol{\theta}_\mathrm{S}\right) = \sum_{n',m'} \sum_{n,m} Y_{n'}^{m'}(\boldsymbol{\theta}_\mathrm{R})\, h_{nm}^{n'm'}(t)\, Y_n^m(\boldsymbol{\theta}_\mathrm{S}), \quad \text{with } h_{nm}^{n'm'}(t) = \sum_i \frac{Y_{n'}^{m'}(\boldsymbol{\theta}_{\mathrm{R},i})\, \mathrm{a}_i\, \delta(t-\tau_i) Y_n^m(\boldsymbol{\theta}_{\mathrm{S},i})}{\mathrm{r}_i}, \tag{3}$$

where $Y_n^m(\boldsymbol{\theta})$ are the SH of order n and degree m, and the expression $h_{nm}^{n'm'}(t)$ denotes a modeled source-and-receiver-directional room impulse response in Ambisonics (SRD ARIR), which we actually measure later on (see Section 2.2).

The directivities $g_A(\boldsymbol{\theta}_A)$, $\mathrm{A} \in \{\mathrm{S}, \mathrm{R}\}$ can be represented by SH expansion:

$$g_A(\boldsymbol{\theta}_\mathrm{A}) = \sum_{n=0}^{N_\mathrm{A}} \sum_{m=-n}^{n} \gamma_{nm}^\mathrm{A} Y_n^m(\boldsymbol{\theta}_A), \tag{4}$$

where N' and N are the maximum orders used to represent the receiver and source directivity, respectively. By inserting Equations (3) and (4) into Equation (2), the integrals in Equation (2) invoke the orthogonality property $\int_{\mathbb{S}^2} Y_n^m(\boldsymbol{\theta}_A) Y_{n'}^{m'*}(\boldsymbol{\theta}_A) \mathrm{d}\boldsymbol{\theta}_A = \delta_{nn'}^{mm'}$ for both source and receiver, yielding a neat sum for the RIR

$$h(t) = \sum_{n'=0}^{N'} \sum_{m'=-n'}^{n'} \sum_{n=0}^{N} \sum_{m=-n}^{n} \gamma_{n'm'}^\mathrm{R} h_{nm}^{n'm'}(t) \gamma_{nm}^\mathrm{S}. \tag{5}$$

For natural, frequency-dependent directivities, multiplication by the spherical-harmonic coefficients of the source and receiver directivity γ_{nm}^S and $\gamma_{n'm'}^\mathrm{R}$ is replaced by convolution with the coefficients of their directional impulse responses $\gamma_{nm}^\mathrm{S} \rightarrow *\gamma_{nm}^\mathrm{S}(t)$ and $\gamma_{n'm'}^\mathrm{R} \rightarrow \gamma_{n'm'}^\mathrm{R}(t)*$, now in the SH domain.

2.2. Measuring the SRD ARIR: Proposed Method

This section presents the proposed efficient SRD ARIR measurement and postprocessing in detail.

Measuring the MIMO RIRs: Here the multiple input multiple output (MIMO) RIRs are measured between a 6-channel compact spherical loudspeaker array (Cubelet) with a radius of 7.5 cm and the 4-channel B-format microphone array (ST450), see Figure 2. The loudspeaker array is equipped with Fountek FR58EX drivers (2 inch coil diameter with $\pm$3 mm maximum linear excursion). A more detailed description on the used arrays and high-resolution directivity measurements can be found online (https://phaidra.kug.ac.at/o:104374).

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

Figure 2. Compact arrays that are used to measure the multiple input multiple output (MIMO) room impulse responses (RIRs). (**a**) Cubelet: Spherical ($r = 7.5\,\text{cm}$) 6-channel loudspeaker array prototype with loudspeakers arranged on surfaces of a cube. (**b**) TSL ST450: 4-channel Ambisonic B-format microphone array with $r = 2\,\text{cm}$.

Omni to omni RIR: Depending on the array geometries, an approximation of the point-to-point omnidirectional RIR $h_0(t)$ can be obtained by transforming both sides of the MIMO RIRs in the SH-domain and extracting the response between the 0th order components. If the array elements are arranged according to a spherical t-design, an approximate of $h_0(t)$ is obtained summing over all channels in the array domain. Please note that the direct path in $h_0(t)$ is ideally a single impulse. However, due to the non-ideal responses of the loudspeakers and microphones as well as the array geometries, even the direct path will be spread in time. A possible approach for improving the omnidirectional response is outlined in [32], but it is not employed here. A denoising of $h_0(t)$ is optional but recommended when experiencing unrealistic long reverberation times. We suggest a denoising strategy that is similar to [33] and it is derived in the Appendix A.

DOA and DOD estimation: Due to the assumption of a temporally and spatially sparse RIR, we address a direction of arrival (DOA) $\boldsymbol{\theta}_\text{R}(t)$ and direction of departure (DOD) $\boldsymbol{\theta}_\text{S}(t)$ to each discrete time instance t of $h_0(t)$. While due to reciprocity any DOA estimation method, e.g., as summarized in [34], can be employed for both DOA and DOD estimation, we use the pseudo intensity vector approach (PIV) as presented by Jarrett et al. [35] for the DOA and an $\boldsymbol{r}_\text{E}$-vector measure [36] related to the magnitude sensor response (MSR) by Politis et al. [37] for determining the DOD.

The DOAs are calculated for the frequencies between 100 Hz and 2.5 kHz. Here, the upper frequency limit is chosen below the spatial aliasing frequencies $f_a = \frac{c}{2\pi r_{ST450}} \approx 3.6\,\text{kHz}$ for $r_{ST450} = 1.5\,\text{cm}$ (defined by $kr_{ST450} = 1$). For the estimation of the DODs, a less restrictive rule is assumed, as it is less affected by linear interference. Here, the upper frequency limit is $f_a = \frac{c}{\pi r_{cubl.}} \approx 1.4\,\text{kHz}$ for $r_{cubl.} = 7.5\,\text{cm}$ (inter-transducer arc length roughly below half a wavelength $\frac{\pi}{2} r_{cubl.} \leq \frac{c}{2f}$). The low cut at 100 Hz minimizes low-frequency disturbance in both the DOA and DOD estimation, respectively. DODs and DOAs become

$$\boldsymbol{\theta}_\text{S}(t) = \frac{\tilde{\boldsymbol{\theta}}_\text{S}(t)}{\left\|\tilde{\boldsymbol{\theta}}_\text{S}(t)\right\|}, \qquad \text{with } \tilde{\boldsymbol{\theta}}_\text{S}(t) = F_L\left\{ \sum_{p=1}^{P} F_{100-1.4\text{k}}\left\{h_{p,0}(t)\right\}^2 \boldsymbol{\theta}_p \right\}, \tag{6}$$

$$\boldsymbol{\theta}_\text{R}(t) = \frac{\tilde{\boldsymbol{\theta}}_\text{R}(t)}{\left\|\tilde{\boldsymbol{\theta}}_\text{R}(t)\right\|}, \qquad \text{with } \tilde{\boldsymbol{\theta}}_\text{R}(t) = F_L\left\{ F_{100-2.5\text{k}}\left\{ \sum_{p=1}^{P} h_{p,0}(t)\right\} F_{100-2.5\text{k}}\left\{ \sum_{p=1}^{P} \boldsymbol{h}_{p,\text{XYZ}}(t)\right\} \right\}, \tag{7}$$

where $\boldsymbol{\theta}_p$ indicates the direction of the p-th loudspeaker, $P = 6$ is the number of array loudspeakers, $h_{p,0}$ the RIR between the p-th loudspeaker and the W channel of the ST450 array, $\|\cdot\|$ is the norm operator, and $h_{p,\text{XYZ}}$ are the first-order channels of the ST450 microphone array. Both the DOA and DOD are computed using a zero-phase band limitation (e.g., by MATLAB's filtfilt with a 4th-order Butterworth band pass) denoted by $F_{\text{f}_\text{l}-\text{f}_\text{u}}$ and a zero-phase temporal smoothing F_L of the resulting estimates using a moving-average Hann window in the interval $[-L/2; L/2]$ for $L = 32$.

SRD ARIR: From Equation 3, and assuming a single propagation path at a time (i.e., assuming temporal disjointness), a first version of the upscaled SRD ARIR becomes

$$\tilde{h}_{nm}^{n'm'}(t) = Y_{n'}^{m'}[\boldsymbol{\theta}_{\mathrm{R}}(t)]\; h_0(t)\; Y_n^m[\boldsymbol{\theta}_{\mathrm{S}}(t)], \tag{8}$$

where the maximum orders $n \leq N$ and $n' \leq N'$ can be chosen freely. The multiplication of the omnidirectional RIR $h_0(t)$ with the SH representations of $\delta[\boldsymbol{\theta}_{\mathrm{R}} - \boldsymbol{\theta}_{\mathrm{R}}(t)]$ and $\delta[\boldsymbol{\theta}_{\mathrm{S}} - \boldsymbol{\theta}_{\mathrm{S}}(t)]$ directionally sharpens the measured SRD ARIR, accordingly. However, the implicit assumption of disjointness (there being only a single DOA and DOD per time sample) is not necessarily true in the late diffuse part of the response. As a result, the temporal fluctuations of $\boldsymbol{\theta}_{\mathrm{R}}(t)$ and $\boldsymbol{\theta}_{\mathrm{S}}(t)$ cause amplitude modulation that potentially corrupt narrow-band spectral properties in $\tilde{h}_{nm}^{n'm'}(t)$. A typical result thereof is a mixing of the longer low-frequency reverberation tails towards higher frequencies, causing unnaturally long reverberation there [21,38], especially as the orders n, n' increase. We propose a scheme for spectral correction which is similar to the one in [38] but adopted for SRD ARIR processing.

In theory, the expected temporal energy decay in an ideal (isotropic) diffuse field should be identical for any source and receiver of random-energy-efficiency-normalized directivity such as the spherical harmonics; this must hold also after decomposition into frequency bands. However, less restrictively, even in non-isotropic diffuse fields, the expected energy decay is identical for subsets of source and receiver directivities that are (pseudo-)omnidirectional: Formal derivation in [38] showed that quadratic summation across same-order spherical harmonics is omnidirectional. Thus, from Equation 8 and with the Unsöld's Theorem [39] $\sum_m |Y_n^m(\boldsymbol{\theta})|^2 = \frac{2n+1}{4\pi}$ for $\boldsymbol{\theta} \in \mathbb{S}^2$ we obtain consistent powers of processed and original RIRs

$$\sum_{m=-n}^{n} \sum_{m'=-n'}^{n'} \left[h_{nm}^{n'm'}(t)\right]^2 = h_0^2(t) \underbrace{\sum_{m'=-n'}^{n'} \left|Y_{n'}^{m'}[\boldsymbol{\theta}_{\mathrm{R}}(t)]\right|^2}_{\frac{2n'+1}{4\pi}} \underbrace{\sum_{m=-n}^{n} \left|Y_n^m[\boldsymbol{\theta}_{\mathrm{S}}(t)]\right|^2}_{\frac{2n+1}{4\pi}} \tag{9}$$

To moreover enforce the short-term energies in $[h_{nm}^{n'm'}(t)]^2$ to become spectrally consistent with those of $h_0^2(t)$, third-octave filtering is useful, where the bth sub-band signal $F_b\{h_0(t)\}$ with center frequency f_b is obtained from a bank of zero-phase filters F_b that is perfectly reconstructing $h_0(t) = \sum_b F_b\{h_0(t)\}$. For every sub-band b and the orders n, n', an energy decay of the upscaled SRD ARIR $F_b\{\tilde{h}_{nm}^{n'm'}(t)\}$ consistent with the original one of $F_b\{h_0(t)\}$ is enforced by envelope correction

$$w_{n,n'}^b(t) = \sqrt{\frac{(2n+1)(2n'+1)}{16\pi^2}} \sqrt{\frac{F_T\{F_b\{h_0(t)\}^2\}}{\sum_{m=-n}^{n}\sum_{m'=-n'}^{n'} F_T\{F_b\{\tilde{h}_{nm}^{n'm'}(t)\}^2\}}}, \tag{10}$$

$$h_{nm}^{n'm'}(t) = \sum_b F_b\{\tilde{h}_{nm}^{n'm'}(t)\} \cdot w_{nn'}^b(t), \tag{11}$$

where $F_T\{\cdot\}$ denotes temporal averaging with a time constant T (e.g., 46 ms).

The simplified *Matlab* source code of the proposed SRD ARIR method can be found in Appendix B.

3. Listening Experiment—Comparative Study

Due to the well-studied perceptual effects [5,7], its well-defined third-order beamforming [24], and its already available high resolution directivity measurements, see, e.g., in [13], the icosahedral loudspeaker (IKO) is employed as the source with controllable directivity throughout the listening experiment. The experiment itself aimed at evaluating the authenticity and perceived externalized localization achieved with the proposed SRD ARIR method, and to compare it with other auralization techniques. The tested five measurement-based auralization techniques (virtualization of the IKO) are described in Section 3.1. An overview of the design and implementation of the listening experiment is

Appl. Sci. **2020**, *10*, 3747

presented in Section 3.2 and insights on the statistical analysis of ratings and the corresponding results are presented in Section 3.3.

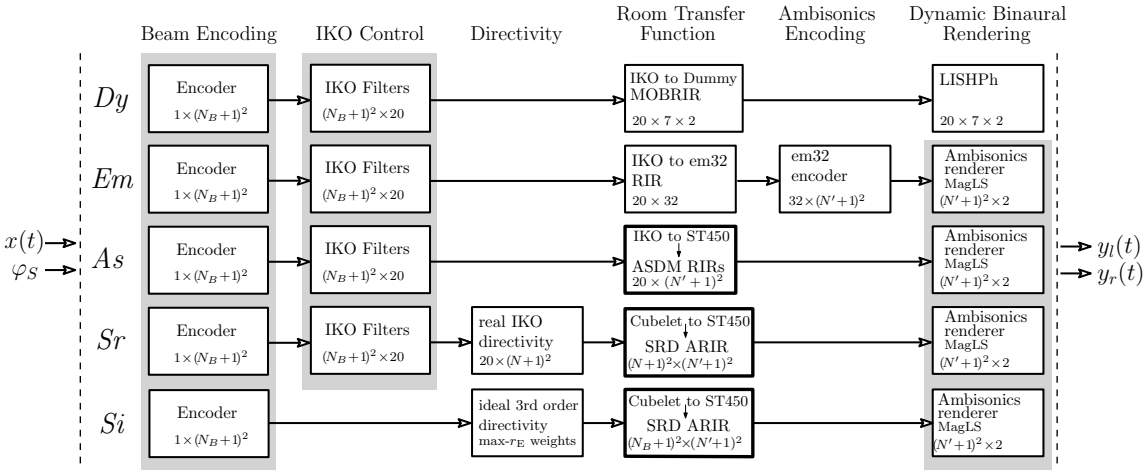


Figure 3. Auralization techniques evaluated in the listening experiment. Here $x(t)$, φ_S, and $y_l(t)$ and $y_r(t)$ are the source signal, the beam direction, and the ear signals of the left and right ear, respectively. The auralization techniques from top to bottom are (i) dummy head BRIR (*Dy*), (ii) MIMO ARIR (*Em*), (iii) Multi ASDM (*As*), (iv) real SRD ARIR (*Sr*), and (v) ideal SRD ARIR (*Si*) auralization. Note that gray-shaded boxes indicate functional blocks which are shared between different techniques. The boxes delimited by bold lines mark techniques in which the actual *Room Transfer Function* is not directly measured but obtained by processing (e.g., upscaling by an ASDM method) of the underlying measurements as proposed in Section 2.2. A detailed description of the techniques can be found throughout Section 3.1.

All underlying measurement data, a short description of the measurement set-up, directivity measurements, and response data as well as the evaluation of the listening experiment are made available online (https://phaidra.kug.ac.at/o:104417).

3.1. Auralization Techniques—Virtualizations of the IKO

As acoustic virtualizations of the IKO, we compared five different auralization techniques in the listening experiment. The block diagram in Figure 3 depicts these techniques and their details are given below. The ear signals $y_l(t)$ and $y_r(t)$ are obtained by running the source signal through several processing stages. Those stages include (i) beam encoding, (ii) IKO control, (iii) directivity, (iv) room transfer function, (v) Ambisonics encoding, and finally (vi) the dynamic binaural rendering.

With the source signal $x(t)$ and the desired beam direction φ_S, the frequency independent encoder outputs the order N_B Ambisonics representation of the beam. The processing in the *Beam Encoding* stage is independent of the auralization technique.

In the *IKO Control* stage the $(N_B + 1)^2$ channels are mapped to 20 loudspeaker signals of the IKO using the frequency-dependent *IKO Filters*. A measurement based approach for designing the multiple-input multiple-output (MIMO) IKO control filters is presented in [40]. It is based on laser Doppler vibrometry measurements and allows for control of side-lobe suppression and excursion-limiting filter design. The designed beam patterns were verified by far-field extrapolated measurements from a surrounding microphone array. The IKO's beamforming can be analyzed by using the open source tool `balloon_holo`, which is part of IEM's Open Data Project (https://opendata.iem.at/projects/dirpat/). All underlying measurements (laser Doppler vibrometry and pressure of a surrounding microphone array) as well as the corresponding *IKO Filters* can be

found online (https://phaidra.kug.ac.at/o:67609) and a summary is presented in [13]; here, we used the IEM IKO3 (https://phaidra.kug.ac.at/o:75316).

In the latest stage, *Dynamic Binaural Rendering* of the Ambisonic scene is obtained by a convolution of the rotated Ambisonic signals with any state-of-the-art FIR binaural Ambisonic renderer. Here, we employ the time-invariant filters of the MagLS (magnitude-least-squares) renderer (The MagLS renderer is part of the IEM plugin suite which can be found here https://plugins.iem.at/) defined in [18,41] to get high-quality Ambisonic rendering already with an order as low as $N = 3$. The perceptual quality improvement of these filters is achieved by using a magnitude-least-squares optimization that disregards phase match in favor of an improved HRTF magnitude at high frequencies. MagLS as outlined in [18,36] also includes an interaural covariance correction that offers an optimal compromise for consistently rendering diffuse fields.

All other processing stages are rather specific per auralization technique and are therefore described separately below.

Dummy head BRIR-based (Dy): The *Directivity* and the *Room Transfer Function* are inherent in the directly measured multiple orientation BRIRs (MOBRIRs) between each loudspeaker of the IKO and the KU100 (https://en-de.neumann.com/ku-100) dummy head. Here, we used an orientation resolution of $\Delta\varphi = 15°$ on an interval between $\varphi = [-45°, \ldots, 45°]$ to obtain the MOBRIRs and the data is available online (https://phaidra.kug.ac.at/o:104386). The *Dy* technique, the *Dynamic Binaural Rendering* is achieved by the linear interpolation with switched high frequency phase (LISHPh) method as described in [42]. In accordance with the findings in [22], setting the crossover $f_c = 2kHz$, $\Delta\varphi = 15°$, and $L = 16$ allows for high-quality BRIR-based binaural rendering, and thus this condition is used as a perceptual target in the study. The processing steps of the reference auralization are shown in the top row of the block diagram in Figure 3. Although the auralization quality (audio quality and spatial mapping) of the *Dy* technique is expected to be high, the measurement effort for the multiple orientations is somewhat enlarged, and the specific dummy head HRIRs cannot be exchanged unless multi-orientation measurement are repeated with other receivers, dummy heads, or individual subjects, separately.

IKO to em32 MIMO RIR (Em): Here, the *Directivity* and the *Room Transfer Function* are represented by the measured array domain MIMO RIRs between the 20 IKO loudspeakers and the 32 microphones of the em32 (https://mhacoustics.com/products). The resulting em32 signals are transformed in the Ambisonics domain using the state-of-the-art encoder presented in [36,40] and are finally binaurally rendered. An evaluation of this specific auralization technique can be found in [25] and the inherent processing stages are depicted in the second row of Figure 3. The underlying MIMO RIRs are accessible online (https://phaidra.kug.ac.at/o:104385). Measuring with the em32 or other higher-order compact spherical microphone arrays increases the hardware effort in terms of channel counts, but permits modular exchange of the receiver directivities or HRIRs, and achieves a native higher-order resolution at the receiver side. Further resolution enhancement by HOSIRR [43] is thinkable but was not used here.

Multi ASDM RIRs (As): This approach employs the first-order tetrahedral ST450 microphone array at the receiver side for measuring the 20×4 (IKO to ST450) MIMO RIRs, which are available online (https://phaidra.kug.ac.at/o:104384). However, the MIMO RIRs are not used directly as the representation of the *Directivity* and *Room Transfer Function*. In a processing stage, the Ambisonic Spatial Decomposition Method (ASDM) [22] is applied to every transducer of the source array; here, the IKO, and the resulting upscaled ASDM RIRs, are eventually used for auralization. This permits a modular exchange of the receiver directivity or HRIRs while keeping the hardware effort at the receiver side minimal. Note that the multi ASDM method is a special form of the SRD ARIR approach, cf. assuming a fixed directivity at the source (the individual loudspeaker) and setting $N = 0$ in Equation (5).

SRD ARIR and real IKO (Sr): The SRD ARIR method as proposed in Section 2 only requires first-order loudspeaker and microphone arrays for measuring the *Room Transfer Function*, on the source and receiver sides respectively. Thus, the SRD method is rather hardware efficient with a theoretical

minimum of 4 channels for the source and the receiver. Here we used our 6-channel Cubelet and the tetrahedral ST450 as source and receiver arrays, respectively. Note that the first-order RIR measurements (https://phaidra.kug.ac.at/o:104376) as well as high resolution directivity measurements of the Cubelet are available online (https://phaidra.kug.ac.at/o:104374).

In a next processing step, the resolution is upscaled from first order to any higher order, see Equation 8 and the detailed description throughout Section 2.2. Therefore, both the source and receiver side are modular and permit exchange with any directivity pattern. In the experiment we inserted KU100 HRIRs with 5th order resolution of a MagLS decoder [18], and at the source side the true measured directivity of the IKO are used. The *Directivity* of the 20 loudspeakers is represented using an order N representation of the directional IRs from every loudspeaker to every microphone of a surrounding microphone array. We use IRs measured using an equiangular grid of 18×36 zenith and azimuth angles, respectively. With 648 sampling points on the sphere we set $N \leq 17$. The high resolution directional IRs of the IKO are available online (https://phaidra.kug.ac.at/o:75316).

SRD ARIR and ideal 3rd-order directivity (Si): While the *Room Transfer Function* is represented by a SRD ARIR as well (same as for *Sr*), the source Directivity is assumed to be an ideal 3rd-order directivity instead of the real IKO, here. Thus, the directivity is synthesized by multiplying the encoded signals with a frequency-independent diagonal matrix containing the max-r_{E} weights [44,45] up to order N_B.

3.2. Design and Implementation

Measurements: The underlying measurements are done in the György Ligeti Saal ($V = 5630\ \mathrm{m}^3$, $T_{60} = 1.4$ s) in Graz, Austria. Figure 4 shows a panoramic photo of the measurement setup and Figure 5 the layout of source, receiver, and the locations of the four reflecting baffle (0.9×1.8 m) positions. Source and receiver were aligned quasi-parallel to the shorter side walls of the room, are facing each other, and are 4.2 m apart. The source–receiver distance approximately corresponds to the critical distance ($r_H = 3.6$ m) when assuming an omnidirectional source, and thus is considered generally interesting. As test signal we used interleaved and exponentially swept sines with a length of 4 s. The measured source and receiver configurations included (i) MOBRIRs (https://phaidra.kug.ac.at/o:104386) between the IKO and multiple dummy head orientations (measurements for *Dy*), (ii) MIMO RIRs (https://phaidra.kug.ac.at/o:104385) between the IKO and the em32 (measurements for *Em*), (iii) MIMO RIRs (https://phaidra.kug.ac.at/o:104384) between the IKO and ST450 (measurements for *As*), and (iv) MIMO RIRs (https://phaidra.kug.ac.at/o:104376) between the Cubelet and the ST450 (measurements for *Sr*, and *Si*).

Figure 4. Panoramic image (360°) of the measurement setup in the György Ligeti Saal, Graz. The camera perspective corresponds with the receiver/listener position.

Appl. Sci. **2020**, *10*, 3747

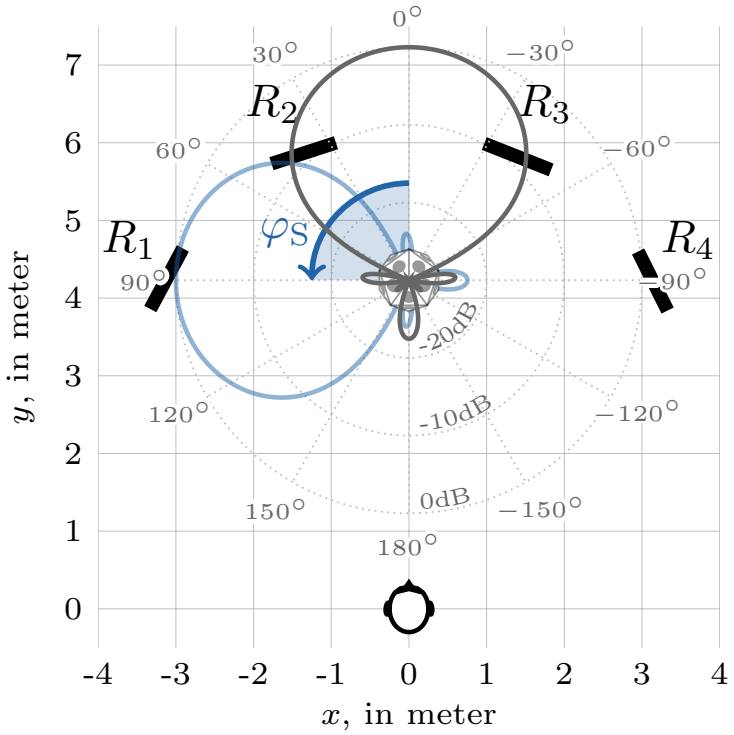


Figure 5. Position of source, listener, and reflectors (R1,..,R4) in György Ligeti Saal, Graz. An ideal 3rd-order max-r_E weighted beampattern with a dynamic range of 30dB is shown in gray. For the listening experiment we used the beam directions $\varphi_S = [0°, 180°, 90°, 45°, -36°, -82°]$.

Tested Directivities: Here, we tested for six distinct beam directions on the horizon $\varphi_S = [0°, 180°, 90°, 45°, -36°, -82°]$, which roughly correspond to the directions of the back wall, the listener, and the reflectors R1, R2, R3, R4, respectively, cf. Figure 5. Those specific directions were chosen as they evoke a pronounced direct or reflection path and allow for testing certain aspects of the auralization:

- 0°: weak direct path, low direct-to-reverberant energy ratio (DRR).
- $-36°, -82°, 45°, 90°$: pronounced reflections and weak direct path. Note that the directions $\varphi_S = -36°$ and $\varphi_S = -82°$ were perceptually chosen by the authors such that the perception of the baffle reflection (R3, and R4) is most pronounced. In order to avoid redundancy due to symmetry and to test for various reflection vs. direct path levels we also included $\varphi_S = 45°$ and $\varphi_S = 90°$ as possible beam directions.
- 180°: pronounced direct path, high direct-to-reverberant energy ratio (DRR).

Implementation: For the sake of reproducibility and in order to circumvent room divergence [46–48], i.e., violation of acoustical expectations arising from the environment in which one listens to headphones, the entire scenery was modeled in virtual reality. In order to deliver graphics as realistically as possible, the room was modeled based on building plans and photogrammetry. Control buttons and labels were added to the virtual environment to give the participants control over the progression of the experimental trials and means to comparatively rate their auditory localization under the various conditions. A screenshot of the user interface is depicted in Figure 6. In addition to the typical playback and save/proceed (upwards facing arrow) buttons we used five colored squares and the corresponding spheres for controlling the experiment. Depending on the tested multistimulus set, those colored squares correspond either to all auralization techniques for a fixed beam direction or to all beam directions for a fixed auralization technique.

Figure 6. Screen shot of VR-rendered environment. Listener faces icosahedral loudspeaker (IKO) and uses controller to switch between conditions when moved over colored floor panels and to drag-and-drop place correspondingly colored auditory event markers (opaque/translucent for active/inactive).

As VR game engine we used Unity (https://unity.com/) and the experimental game was played using the HTC VIVE, i.e., system comprising head-mounted display (HMD), controllers, and tracking. The tracking data, i.e., head rotations, from the HMD was sent to Reaper (https://www.reaper.fm) via OSC [49], where the audio processing was implemented.

While Rakerd and Hartmann [50] stated that short onsets and transient signals overall simplify localization, Wendt et al. [7] discovered that such signals are localized significantly closer to the IKO. In order to create a large scenery of perceivable auditory objects distributing to various remote locations with regard to the IKO, Wendt et al. [7] recommends using signals with slow onset. For conditions with clear effects, we therefore used a 1.5 s long pink noise burst with fade-in and fade-out times of 500 ms (linear fades) and 500 ms silence at the end.

For encoding and multi-channel convolution with IKO control filters, directivities, and RIRs we used the `mcfx` (http://www.matthiaskronlachner.com/?p=1910) plug-ins and as binaural renderer of the Ambisonic signals we used the `BinauralDecoder` (https://plugins.iem.at/) [18,41]. The ear signals were played back via headphones (AKG 702) plugged into an external audio interface (RME MadiFace & RME FireFace UCX). Note that an orientation mismatch $< 5°$ between different arrays used for measuring the RIRs (cf. the IKO vs. Cubelet, and ST450 vs. em32) can almost not be avoided. Thus, the authors perceptually aligned the auralization techniques for $\varphi_S = 180°$.

During informal listening experiments (by four participants) we found that all auralization techniques under test obey a high overall sound quality (no artifacts or temporal smearing). However, the overall timbre slightly varies across the techniques as we employed technique specific measurement hardware (e.g., Cubelet vs. IKO). While a global and steering-direction-independent equalization was not feasible, the techniques were perceptually equalized using a parametric multi-band EQ for a fixed steering direction $\varphi_S = 180°$ (pointing to the listener).

Input Method: During the experiment, participants were asked to indicate the position (i.e., direction and distance relative from the listener) of the perceived sound and to follow a certain procedure: (i) point to a colored square to select a stimulus for looped playback, (ii) pick-up the correspondingly colored ball by pointing towards it and pressing the trigger, (iii) with trigger pressed, point to the perceived direction and adjust the distance by moving the thumb on the controller track pad, (iv) release the trigger to drop off the ball at the intended position, (v) proceed until all balls are positioned, then save responses and proceed to next multistimulus set. Participants were allowed to reposition any ball as often as desired until the responses of the entire multistimulus set were logged in.

Design: The experiment consisted of 12 multistimulus sets, of which the first one was part of a training and familiarization phase. In the following 11 sets, participants were asked to rate 5 stimuli per set. Those 5 stimuli either consisted of all the five auralization techniques and a fixed beam direction (6 sets), or of all beam directions (except the $-45°$ beam direction) (In order to keep only 5 stimuli per scene, scenes with a given auralization technique were not containing the $-45°$ beam direction. For this reason results obtained for the $-45°$ beam direction were not used in the statistical analysis.) for a fixed auralization technique (5 sets). Both the order of sets, as well as the assignment of a stimulus to a certain colored square within the set were randomized. The 13 participants (normal hearing, all male, age between 24–52) were asked to repeat the experiment in order to provide a second response per set. Correspondingly, most of the 13 participants (except 1 who did not repeat the experiment) evaluated $11 \cdot 5 \cdot 2 = 110$ stimuli.

3.3. Results and Discussion

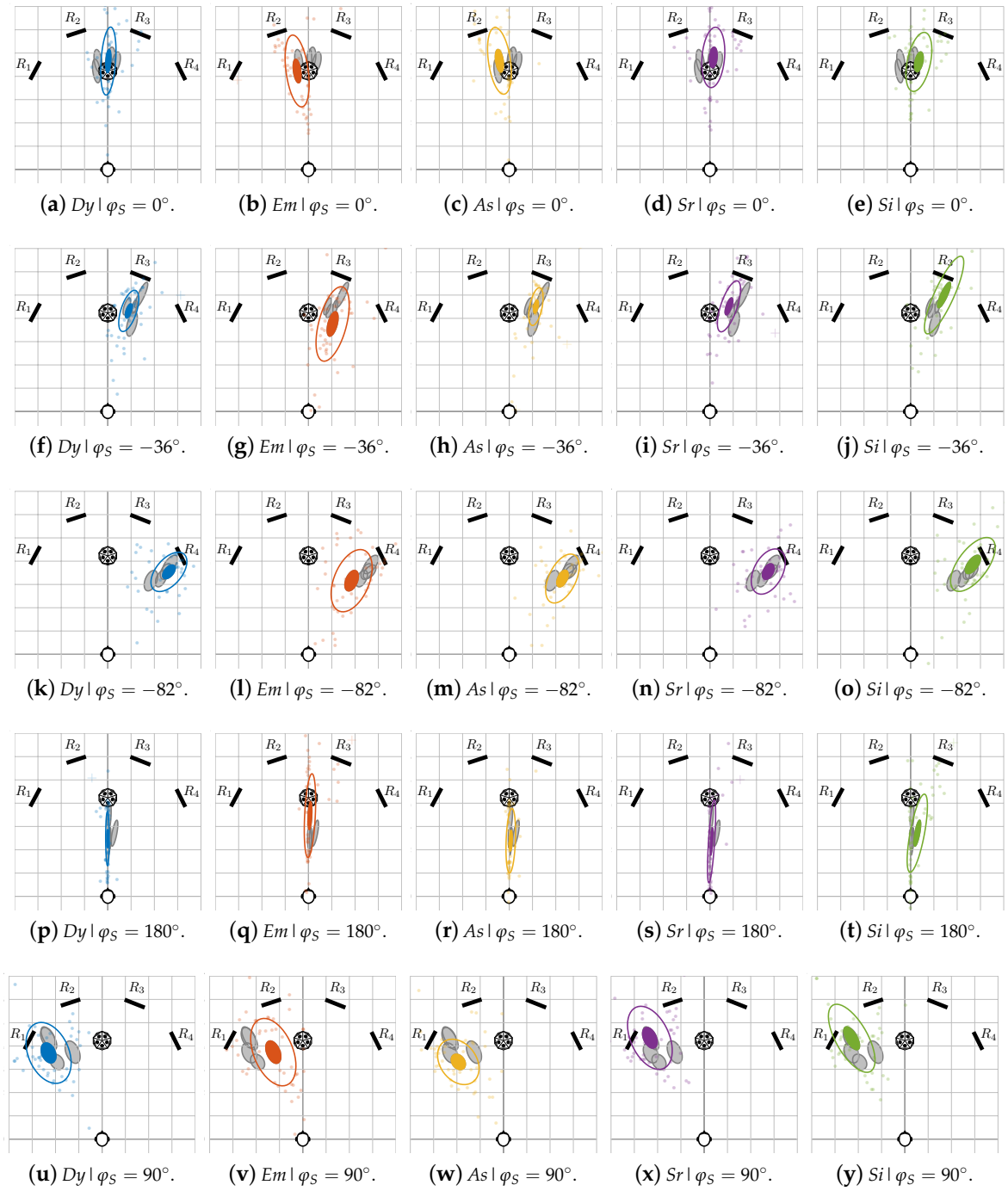
The positions of the perceived sound objects are given in the Cartesian coordinate system with the listener at the origin. As results show little to no variation in height (z coordinate), we focus on an evaluation of the x, y coordinates.

Overall inspection in two dimensions: In a first processing step, outliers are defined as responses lying outside a Mahalanobis distance (in estimated standard deviations) of three within a preliminary, non-robust analysis. After removal of the outliers, we use bivariate statistical analysis to estimate the means and their standard deviation and 95% confidence region according to Hotelling's T2 distribution (see [51] Ch. 3). The result of this analysis are depicted in Figure 7, where data points, outliers, and standard deviation and confidence region ellipses are indicated as dots, crosses, and not-filled and filled ellipses, respectively. In case of similar sizes of the statistical spreads, statistically significant differences may be inspected by observing whether the mean value of one condition lies outside the 95% ($p < 0.05$) confidence ellipses of the other conditions.

While each row of Figure 7 depicts the results for all auralization techniques and a certain beam direction φ_S, each column shows the perceived position of the auditory events per technique and for all beam directions $\varphi_S = [0°, -36°, -82°, 180°, 90°]$. Thus, comparison within the rows is used to identify differences across auralization techniques, and comparison within each column gives indication if each auralization technique is able to reproduce the well-described perceptual effects of the IKO [7], or similar devices. These effects are explained by exciting pronounced propagation paths and dimming the direct path, which is known to evoke auditory events whose position needs not coincide with the physical source. Moreover, Wendt et al. [5] showed that the IKO's directivity allows for altering the DRR and thus, for controlling the perceived distance, e.g., by steering the beam towards or away from the listener.

Overall, we observe that all techniques are qualitatively able reproduce the perceptual effects known from studies involving the physical IKO [5,7], cf. columns in Figure 7. The ratings show a clear consensus with the expected positions of the auditory events, i.e., by steering a beam towards a reflector $\varphi_S = [-36°, -82°, 90°]$ the auditory event is located near the respective reflector baffle. Moreover, steering the beam towards $\varphi_S = 0°$ away from the listener and $\varphi_S = -180°$ towards the listener either evokes an auditory event behind the physical IKO or a very close one, respectively. We found that the ratings per beam direction φ_S are significantly different for all auralization techniques.

A detailed analysis of the differences related to the auralization techniques, cf. rows in Figure 7, is done for independent univariate attributes. These univariate attributes were not asked separately in the listening experiment, but they are obtained for the subsequent analysis by mapping of the responses to the following independent attributes (i) localizability, (ii) the direction, and (iii) distance. This analysis is based on the following considerations.



(a) $Dy \mid \varphi_S = 0°$. **(b)** $Em \mid \varphi_S = 0°$. **(c)** $As \mid \varphi_S = 0°$. **(d)** $Sr \mid \varphi_S = 0°$. **(e)** $Si \mid \varphi_S = 0°$.

(f) $Dy \mid \varphi_S = -36°$. **(g)** $Em \mid \varphi_S = -36°$. **(h)** $As \mid \varphi_S = -36°$. **(i)** $Sr \mid \varphi_S = -36°$. **(j)** $Si \mid \varphi_S = -36°$.

(k) $Dy \mid \varphi_S = -82°$. **(l)** $Em \mid \varphi_S = -82°$. **(m)** $As \mid \varphi_S = -82°$. **(n)** $Sr \mid \varphi_S = -82°$. **(o)** $Si \mid \varphi_S = -82°$.

(p) $Dy \mid \varphi_S = 180°$. **(q)** $Em \mid \varphi_S = 180°$. **(r)** $As \mid \varphi_S = 180°$. **(s)** $Sr \mid \varphi_S = 180°$. **(t)** $Si \mid \varphi_S = 180°$.

(u) $Dy \mid \varphi_S = 90°$. **(v)** $Em \mid \varphi_S = 90°$. **(w)** $As \mid \varphi_S = 90°$. **(x)** $Sr \mid \varphi_S = 90°$. **(y)** $Si \mid \varphi_S = 90°$.

Figure 7. Bivariate statistical analysis of the perceived position of the auditory event per beam direction φ_S (columns) and per auralization technique (rows). Data points, outliers, and standard deviation and 95% confidence region ellipses are indicated as dots, crosses, and not-filled and filled ellipses, respectively.

As defined by Lindau et al. [52], localizability is related to the ability to asses the spatial extent and location of a sound source. If this task is difficult, the localizability is low and if localizability is high, a sound source is clearly delimited. Moreover, localizability is often associated with the perceived extent of as sound source and thus we assume that the area of the standard deviation ellipse can be used as an indication of the localizability.

Appl. Sci. **2020**, *10*, 3747

The two-dimensional source position indications yield a clear bivariate distribution, cf. Figure 7, and with a mean angular offset between the main axis of the standard deviation ellipse and the mean perceived direction φ_p (defined by listener and mean position of the perceived sound) of only 3.52°, we may assume the variations to be independent along the perceptual axis of distance and direction.

As this visual evaluation may be difficult, we use a Wilcoxon signed-rank test [53] with a Bonferroni–Holm correction [54] to determine p-values of pairwise comparisons between test conditions and define $p < 0.05$ as significantly different throughout this article. We employ nonparametric statistics as we do not assume a normal distribution of ratings and due to the correction (David Groppe, 2020, Bonferroni–Holm Correction for Multiple Comparisons, https://www.mathworks.com/matlabcentral/fileexchange/28303-bonferroni-holm-correction-for-multiple-comparisons) p-values can exceed the expected range and thus $p > 1$ is valid. The *Matlab* script of the statistical analysis and the raw listener ratings are available online in the accompanying project (https://phaidra.kug.ac.at/o:104416). A detailed discussion of the results in terms of localizability, direction, and distance is given below.

Localizability: The median values and 95% confidence interval of the area under the standard deviation ellipse pooled for all beam directions φ_S are depicted per auralization technique in Figure 8. While the median values indicate highest and lowest localizability for the *Dy* and *Em* techniques, respectively, the differences among all techniques is not significant ($p > 1.7$).

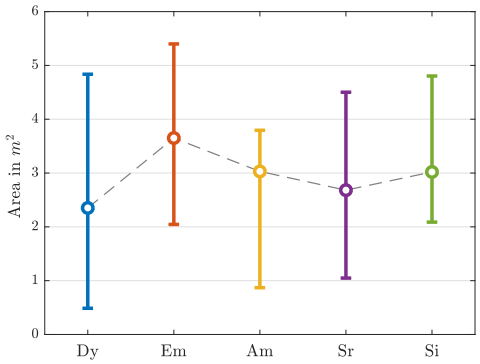
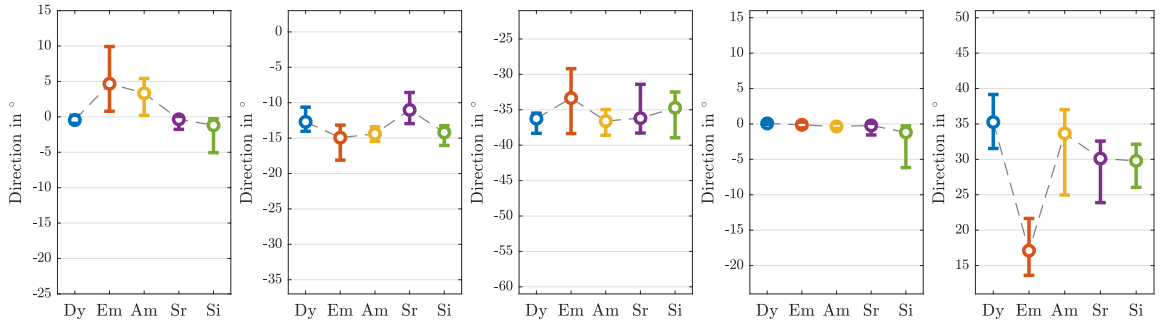


Figure 8. Median value and 95% confidence interval of the area under the standard deviation ellipse pooled for all beam directions φ_S.

Direction: All ratings are transformed into a polar coordinate system with the listener at the center and are analyzed for azimuth and radius, i.e., direction and distance, separately. Due to the findings in [22] we assume that the *Dy* auralization can be used as the reference condition. Thus, the p-values are given for testing the significance levels between *Dy* and all other conditions. The median values and 95% confidence intervals for all beam directions and auralization techniques are shown in Figure 9 and the in p-values between the reference condition (*Dy*) and the other techniques are presented in Table 1.

Appl. Sci. **2020**, *10*, 3747



(a) $\varphi_S = 0°$. **(b)** $\varphi_S = -36°$. **(c)** $\varphi_S = -82°$. **(d)** $\varphi_S = 180°$. **(e)** $\varphi_S = 90°$.

Figure 9. Median value and 95% confidence interval of direction ratings per beam direction φ_S. Please note the varying *y*-axis as per sub-figure a range of $[15°, -25°]$ around the median of the *Dy* ratings is shown.

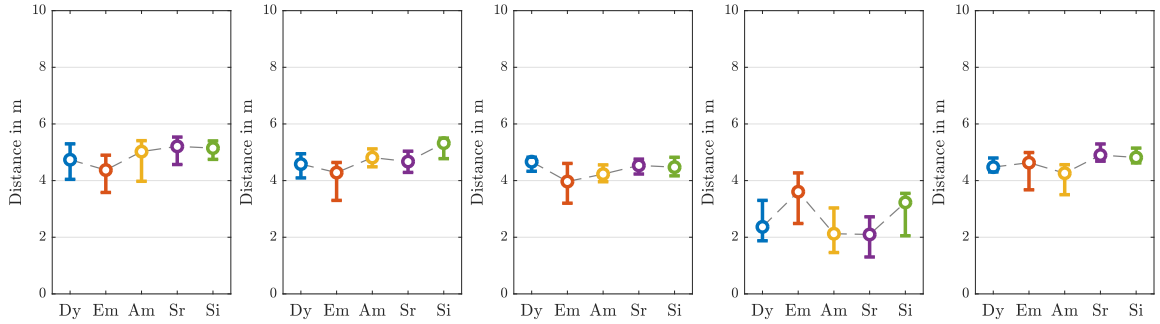
Table 1. *p*-values (Wilcoxon signed-rank test with Bonferroni–Holm correction) for ratings of direction. Insignificant differences (*p*-values ≥ 0.05) are indicated by bold numbers.

	$\varphi_S = 0°$	$\varphi_S = -36°$	$\varphi_S = -82°$	$\varphi_S = 180°$	$\varphi_S = 90°$
			Dy		
Em	0.00	**0.15**	**1.38**	**0.70**	0.00
As	0.00	**0.08**	**1.77**	**0.17**	**0.52**
Sr	**0.76**	**0.23**	**3.76**	**0.07**	**0.07**
Si	**0.30**	**0.15**	**3.76**	0.00	0.03

Despite an almost symmetrical set-up (cf. Figure 5) and only a slight change in beam direction, the direction results for the beams steered towards $\varphi_S = -82°$ and $\varphi_S = 90°$ show some deviation. Ratings for $\varphi_S = 90°$ are less consistent (larger confidence interval) and the auditory event is localized more closely to the IKO for all techniques, when compared to $\varphi_S = -82°$. This can be explained when taking a closer look at the (ideal) 3rd-order max-r_E weighted beam pattern as depicted in Figure 5. While for $\varphi_S = 90°$ a side lobe is pointing towards the listener, this side lobe is almost avoided (-5dB lower) for $\varphi_S = -82°$.

Overall, there is no significant difference between the directional mapping of the *Dy* and *Sr* techniques. For all other techniques we found significant differences for some beam directions. The *Em*-based auralization was the least consistent and produced the smallest lateralisation for $\varphi_S = -82°$ and $\varphi_S = 90°$ compared to the other techniques. This is particularly pronounced for $\varphi_S = 90°$ where there is more direct sound.

Distance: The median values and 95% confidence intervals for all beam directions and auralization techniques are shown in Figure 10 and the in *p*-values between the reference condition (*Dy*) and the other techniques are presented in Table 2.



(a) $\varphi_S = 0°$. **(b)** $\varphi_S = -36°$. **(c)** $\varphi_S = -82°$. **(d)** $\varphi_S = 180°$. **(e)** $\varphi_S = 90°$.

Figure 10. Median value and 95% confidence interval of distance ratings per beam direction φ_S. Please note the fixed y-axis showing the interval of $[0, 10]$m form the listener.

We found that the congruence of the distance mapping is high for all tested auralization techniques, almost independently of the specific beam directions φ_S. The only exception is the *Em* based auralization, where for $\varphi_S = 180°$ the source is perceived significantly further away.

Table 2. *p*-values (Wilcoxon signed-rank test with Bonferroni–Holm correction) for ratings of distance. Insignificant differences (p-values ≥ 0.05) are indicated by bold numbers.

	$\varphi_S = 0°$	$\varphi_S = -36°$	$\varphi_S = -82°$	$\varphi_S = 180°$	$\varphi_S = 90°$
			Dy		
Em	**1.14**	**0.78**	**0.49**	0.02	**1.27**
As	**3.40**	**0.78**	**0.91**	**1.52**	**0.47**
Sr	**3.24**	**0.65**	**2.21**	**1.53**	**0.37**
Si	**3.40**	0.04	**1.52**	**1.42**	**0.47**

Discussion: The five different auralization techniques used to virtualize the IKO all involve measurements of RIRs. The results from the presented listening experiment verify that all tested techniques are able to qualitatively reproduce the perceptual effects known from studies involving the physical IKO [5,7], cf. ellipses in Figure 7. Moreover, still, a detailed analysis of the perceptual attributes of direction, and distance, indicates some significant differences to the reference condition (dummy head based rendering *Dy*) for specific combinations of technique and beam direction, with some noticeable trend.

In order to give an indication of the directional mapping quality of all tested auralization techniques, the mean direction offset to the *Dy* technique is listed in Table 3. Overall, the *Em* techniques yields with 5.72° the largest incongruence in directional mapping, while the errors of *As*, *Sr*, and *Si* roughly stay below 2°.

Table 3. Mean direction offset to dy per auralization technique over all tested beam directions φ_S.

Em	***As***	***Sr***	***Si***
5.72°	1.57°	1.46°	2.01°

The mean distance offset between *Dy* and all other techniques pooled for all beam directions φ_S is given in Table 4. With a mean distance error of 0.28m the *As*, and *Sr* auralizations clearly outperform the *Em*, and *Si* techniques.

Appl. Sci. **2020**, *10*, 3747

Table 4. Mean absolute distance offset to *Dy* per auralization technique over all tested beam directions φ_S.

Em	*As*	*Sr*	*Si*
0.55 m	0.28 m	0.28 m	0.51 m

Although there is no clear perceptual winner, it still appears that the *As* and *Sr* approaches match the *Dy* reference best. We assume that using the first-order source (Cubelet) enhances the flexibility as its frequency range for directivity synthesis is larger than the one of the IKO because of its smaller size. Its upscaling permits a modular exchange of the directivity to arbitrary artificial or measured higher-order directivity patterns. On the receiver side, measurements with the dummy head (*Dy*) are not modular in terms of exchanging the receiver directivity patterns in terms of other HRIRs, while the measurements with the em32 and ST450 microphone arrays permit HRIR exchange. The em32 reaches a native resolution up to the 4th order, whereas the 1st-order ST450 is less demanding in terms of channel count, has no satisfactory native resolution but allows to be upscaled to higher orders.

Our impression is that auralization involving a first-order receiver and either the highly directional source or its first-order source as a replacement tend to work most reliably. Essentially, the results allow us to recommend the SRD ARIR (*Sr*) model and processing method for its high degree of modularity and reduction of measurement hardware and effort. It reaches perceptual qualities comparable to rendering based on dummy-head measurements (*Dy*), while higher-order directional sources can be exchangeably interfaced with the processed and upscaled SRD ARIRs. It is necessary to mention that the particular prototype employed as first-order measurement source (Cubelet) in our study is not necessarily powerful enough for every application, for instance, when the signal-to-noise ratio is low because of background noise. In such cases, stronger alternatives could be considered [55].

4. Conclusions

In this contribution we presented the concept and a comparative perceptual evaluation of a source-and-receiver-directional ARIR capture and processing approach (SRD ARIR) with a variety of technical alternatives. Although the proposed SRD ARIR rendering method only employs a small set of first-order directivities (omnidirectional and figures of eight aligned with x, y, and z) in the measurement, the approach produced auralization of higher-order directional source and receiver configurations that was performing well in the comparison. Its directional resolution enhancement involves the Ambisonic spatial decomposition method (ASDM) that we could extend to both sides of the measured ARIRs.

In the dynamic headphone-rendering-based evaluation, we employed the highly directional icosahedral loudspeaker (IKO) as a virtual source because of its well-described measured directivity and the well-studied perceptual effects it causes. For the sake of reproducibility and to obey optimal externalization the auralization listening experiment was done within a head-mounted-display visualization of the virtual environment. Interactive visual objects were used to indicate auditory event locations in space.

The proposed SRD ARIR method performed similarly accurate as the reference auralization based on multiple-orientation binaural room impulse responses (MOBRIRs). We found no significant difference for the perceptual attributes of localizability, direction, and distance. Although most of the alternative techniques performed comparable to the reference auralization, the SRD ARIR technique has benefits in terms of modularity and efficiency: It only requires a small number of hardware channels, and SRD ARIRs offer a generic interface between RIRs and source and receiver directivities. Any application requiring a flexible exchange of directivities can potentially benefit from the small number of responses needed to characterize the room (be it measured or simulated).

A collection of room responses measured for the study, responses of the listening experiment, the statistical analysis, and high-resolution directivities of the arrays is made available online (https://phaidra.kug.ac.at/o:104417).

Author Contributions: Conceptualization, M.Z. and F.Zo.; Writing—Original Draft Preparation M.Z. and F.Zo with periodic contributions by F.Za.; Writing—Review & Editing, M.Z. with contributions of F.Zo., and F.Za.; Software, M.Z. with periodic contributions by F.Zo and F.Za; Listening Experiment Implementation, M.Z.; Listening Experiment Design, M.Z., F.Zo., and F.Za.; Listening Experiment Conduction: F.Za with periodic contributions by M.Z.; Data Analysis, F.Za with periodic contributions by F.Zo. and M.Z.; Measurements, M.Z., F.Zo., and F.Za. All authors have read and agreed to the published version of the manuscript.

Funding: The major part of the work was carried out within the OSIL project AR 328-G21 that was funded by the Open Access Funding by the Austrian Science Fund (FWF).

Acknowledgments: We thank Matthias Frank for his companionship in pursuing the proposed technique over the past years, and all voluntary listeners for their participation in our listening experiments.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Transition to noise in Schröder backwards integrated impulse response

The late part of a noisy impulse response of a room is well modeled by

$$h(t) = a\, n_1(t)\, e^{-bt} + c\, n_2(t), \tag{A1}$$

where $n_{1,2}(t)$ are two uncorrelated normalized noise processes, $n_1(t)$ as the carrier of the diffuse, decaying reverberation, and $n_2(t)$ as stationary background or measurement noise. The expected squared impulse response is

$$E\{h^2(t)\} = a^2 \underbrace{E\{n_1^2(t)\}}_{=1} e^{-2bt} + c^2 \underbrace{E\{n_2^2(t)\}}_{=1} + 2a^2c^2 \underbrace{E\{n_1(t)n_2(t)\}}_{=0} e^{-bt} = a^2\, e^{-2bt} + c^2. \tag{A2}$$

The expected Schröder backwards integrated impulse response yields

$$\begin{aligned} S(t) &= \int_{\mathrm{T}_{\gg}}^{t} E\{h^2(t)\}\, \mathrm{d}t = \left[\frac{a^2\, e^{-2bt}}{-2b} + c^2 t\right]_{\mathrm{T}_{\gg}}^{t} = \frac{a^2\,(e^{-2bt} - e^{-2b\mathrm{T}_{\gg}})}{2b} + c^2\,(\mathrm{T}_{\gg} - t) \\ &= \frac{a^2\, e^{-2bt}}{2b} + c^2\,(\mathrm{T}_{\gg} - t). \end{aligned} \tag{A3}$$

Within an early section the contribution of the stationary noise amplitude $c^2(T_{\gg} - t)$ is negligible compared to the energy decay $S(t) \to S_{\mathrm{c}=0}(t)$, and linear regression from the observed $\ln S_{\mathrm{c}=0}(t)$ yields the coefficients a and b for the noise-free case by

$$\ln S_{\mathrm{c}=0}(t) = \ln \frac{a^2\, e^{-2bt}}{2b} = -2bt + \ln \frac{a^2}{2b}. \tag{A4}$$

At later time segments, in particular where the model $S_{c\to 0}(t)$ based on a and b is exceeded by the observed $S(t)$, say, by a factor of 10, one can estimate c from linear regression of $S(t) \to S_{a=0}(t)$

$$S_{a=0}(t) = c^2(\mathrm{T}_{\gg} - t). \tag{A5}$$

Together with the known regression parameters b, a, and c, the envelope of the squared impulse response in Equation (A2) can be enforced to take the shape as if c were zero

$$E\{h^2(t)\, w^2(t)\} = E\{h_{c=0}^2(t)\}. \tag{A6}$$

Appl. Sci. **2020**, *10*, 3747

Obviously, the envelope correction that restores the model envelope of a noise-free room impulse response becomes $w^2(t) = \frac{E\{h^2_{c=0}(t)\}}{E\{h^2(t)\}} = \frac{a^2 e^{-bt}}{a^2 e^{-bt} + c^2}$ and therefore

$$w(t) = \frac{1}{\sqrt{1 + \frac{c^2}{a^2}\, e^{2bt}}}. \tag{A7}$$

Appendix B. SRD ARIR

The Matlab source code of the proposed SRD ARIR method, a source-and-receiver-directional version of the Ambisonic Spatial Decomposition Method (ASDM [22]) can be found in Listing 1. It requires the spherical harmonics implemented in Politis' lirbrary (https://github.com/polarch/Spherical-Harmonic-Transform=. The SRD algorithm becomes ASDM when choosing $N_S = 0$ for the source.

```
%% First order B-format RIR (ARIR)
[x,fs] = audioread(fname); % sorted in <W,X,Y,Z>from ls1, <W,X,Y,Z> from ls2, ... until ls6
% fs is the sampling frequency
x = reshape(x,[size(x,1) 4 size(x,2)/4]); % Nsamples x 4(ST450) x 6(c.sph.lspk.)

%% Parameters and Settings
N = 7; % Ambisonic order
Nfft = 2^ceil(log2(size(x,1))); % fft length
Lsmooth_dirfluct = 32;  win_dirfluct=hann(Lsmooth_dirfluct); % smoothing of fluctuating DODs/DOAs
Lsmooth_specdecay = 2048;   win_specdecay=hann(Lsmooth_specdecay); % smoothing for fs = 44.1kHz
Nlate=7000; Nfadelate=1000; win_crossf=sin(pi/2 *(1:Nfadelate)').^2; % fade to spectrally corrected late RIR

%% PIV DOA estimation (for omnidirectional source)
[b,a] = butter(4,[100 2500]/(fs/2)); % bandpass with fl = 100Hz and fh = 1.5kHz
xbp = filtfilt(b,a,sum(x,3));
ix = xbp(:,1).*xbp(:,2);   ix = circshift(fftfilt(win_dirfluct,ix),-floor(Lsmooth_dirfluct/2));
iy = xbp(:,1).*xbp(:,3);   iy = circshift(fftfilt(win_dirfluct,iy),-floor(Lsmooth_dirfluct/2));
iz = xbp(:,1).*xbp(:,4);   iz = circshift(fftfilt(win_dirfluct,iz),-floor(Lsmooth_dirfluct/2));
azi = atan2(iy, ix);    zen = atan2(sqrt(ix.^2+iy.^2),iz);
Ydoa = getSH(N,[azi,zen],'real'); % from https://github.com/polarch/Spherical-Harmonic-Transform/

%% rE-magnitude-based DOD estimation (for omnidirectional receiver)
[b,a] = butter(4,[100 1400]/(fs/2)); % bandpass with fl = 100Hz and fh = 1.4kHz
xbp = filtfilt(b,a,squeeze(x(:,1,:)));
U = [ 1  0  0; 0  1  0; -1  0  0; ... % 6 array loudspeaker direction vectors (transposed)
      0 -1  0; 0  0 -1; 0  0  1];
E = xbp.^2;   E = circshift(fftfilt(win_dirfluct,E),-floor(Lsmooth_dirfluct/2));
rE = E * U;
azi = atan2(rE(:,2), rE(:,1));   zen = atan2(sqrt(sum(rE(:,1:2).^2,2)),rE(:,3));
Ydod = getSH(N,[azi,zen],'real'); % from https://github.com/polarch/Spherical-Harmonic-Transform/

%% SRD Upscaling
x = (sum(x(:,1,:),3) .* Ydoa) .* reshape(Ydod,[size(Ydod,1) 1  size(Ydod,2)]);

%% Spectral Decay Correction
H = thirdoctave_filter_bank_linph(Nfft,fs);
x_c = zeros(Nfft,(N+1)^2, (N+1)^2); % corrected upscaled ARIR
x_c(1:size(x,1),1,1) = x(:,1);
for k = 1:size(H,2)
    xthird0 = ifft(fft(x(:,1,1),Nfft).*H(:,k));
    xthird0rms = sqrt(circshift(fftfilt(win_specdecay,xthird0.^2),-Lsmooth_specdecay/2));
    for nrcv = 1:N
        nidx_rcv = nrcv^2+(1:2*nrcv+1);
        for nsrc = 1:N
            nidx_src = nsrc^2+(1:2*nsrc+1);
            xthirdn = ifft(fft(x(:,nidx_rcv,nidx_src),Nfft).*H(:,k));
            xthirdnrms = sqrt(sum(circshift(fftfilt(win_specdecay,xthirdn(:,:).^2),-Lsmooth_specdecay/2),2));
```

```
            w_c = xthird0rms./(xthirdnrms+1e-6)*sqrt((2*nsrc+1)*(2*nrcv+1)); % correction window
            x_c(:,nidx_rcv,nidx_src) = x_c(:,nidx_rcv,nidx_src)+xthirdn.*w_c;
        end
    end
end
% no spectral correction in early part:
idx_xfade = Nlate+(0:Nfadelate-1)
x(idx_xfade,:) = (1-win_late) .* x(idx_xfade,:,:) + win_late .*x(idx_xfade,:,:);
x(idx_xfade(end)+1:end,:) = xc(idx_xfade(end)+1:end,:); % x contains N3D normalized upmixed SRD RIR

%% DFT-Domain Filter Bank
function H = thirdoctave_filter_bank_linph(Nfft,fs)
f=linspace(0,fs/2,Nfft/2+1);
f(1)=f(2)/4;
fc = 25*2.^(0:1/3:9.9); %third-octave vector
H = zeros(Nfft/2+1,length(fc));
for k = 1:length(fc)
    nthoctaves = log2(f/fc(k))*3; % 3rd-octaves distance from center freq.
    upper = 1.0*(k<length(fc)); % upper 3rd-octave limit (high-pass in last band)
    lower =-1.0*(k>1); % lower 3rd-octave limit (low-pass in first band)
    nthoctaves = max(min(nthoctaves,upper,lower);
    H(:,k) = cos(nthoctaves*pi/2).^2;
end
H = [H;flipud(H(2:end-1,:))];
end
```

Listing 1: MATLAB simplified source code of the proposed SRD ARIR method.

References

1. BS EN ISO 3382-1:2009. Acoustics — Measurement of room acoustic parameters. Part 1:Performance spaces. In *British Standard*; Cambridge University Press: Cambridge, UK, 2009; pp. 1–26.
2. Otondo, F.; Rindel, J.H. The Influence of the Directivity of Musical Instruments in a Room. *Acta Acust. Acust.* **2004**, *90*, 1178–1184.
3. Vigeant, M.C.; Wang, L.M.; Rindel, J.H. Investigations of multi-channel auralization technique for solo instruments and orchestra. In Proceedings of the 19th International Congress on Acoustics, Madrid, Spain, 2–7 September, 2007.
4. Laitinen, M.V.; Politis, A.; Huhtakallio, I.; Pulkki, V. Controlling the perceived distance of an auditory object by manipulation of loudspeaker directivity. *J. Acoust. Soc. Am.* **2015**, *137*, 462–468, doi:10.1121/1.4921678.
5. Wendt, F.; Zotter, F.; Frank, M.; Höldrich, R. Auditory distance control using a variable-directivity loudspeaker. *Appl. Sci.* **2017**, *7*, 666. doi :10.3390/app7070666.
6. Ronsse, L.M.; Wang, L.M. Effects of room size and reverberation, receiver location, and source rotation on acoustical metrics related to source localization. *Acta Acust. Acust.* **2012**, *98*, 768–775, doi:10.3813/AAA.918558.
7. Wendt, F.; Sharma, G.K.; Frank, M.; Zotter, F.; Höldrich, R. Perception of Spatial Sound Phenomena Created by the Icosahedral Loudspeaker. *Comput. Music J.* **2017**, *41*, 76–88, doi:10.1162/COMJ.
8. Wang, L.M.; Vigeant, M.C. Evaluations of output from room acoustic computer modeling and auralization due to different sound source directionalities. *Appl. Acoust.* **2008**, *69*, 1281–1293, doi:10.1016/j.apacoust.2007.09.004.
9. Khaykin, D.; Rafaely, B. Acoustic analysis by spherical microphone array processing of room impulse responses. *J. Acoust. Soc. Am.* **2012**, *132*, 261–270, doi:10.1121/1.4726012.
10. Morgenstern, H.; Klein, J.; Rafaely, B.; Noisternig, M. Experimental investigation of multiple-input multiple-output systems for sound-field analysis. In Proceedings of the 22nd International Congress on Acoustics, Buenos Aires, 5–9 September, 2016.
11. Alary, B.; Massé, P.; Välimäki, V.; Noisternig, M. Assessing the Anisotropic Features of Spatial Impulse Responses. In *Proceedings of the EAA Spatial Audio Signal Processing Symposium*; EAA: Oshkosh, WI, USA, 2019; pp. 43–48, doi :10.25836/sasp.2019.32.

12. Nolan, M.; Verburg, S.A.; Brunskog, J.; Fernandez-Grande, E. Experimental characterization of the sound field in a reverberation room. *J. Acoust. Soc. Am.* **2019**, *145*, 2237–2246, doi:10.1121/1.5096847.
13. Schultz, F.; Zaunschirm, M.; Zotter, F. Directivity and electro-acoustic measurements of the IKO. In Proceedings of the 144th AES Convention, Milan, Italy, 23–26 May 2018.
14. Neal, M.T. A Spherical Microphone and Compact Loudspeaker Array Measurement Database for the Study of Concert Hall Preference. Ph.D. Thesis, The Pennsylvania State University, State College, PA, USA, 2019.
15. Oberem, J.; Richter, J.G.; Setzer, D.; Seibold, J.; Koch, I.; Fels, J. Experiments on localization accuracy with non-individual and individual HRTFs comparing static and dynamic reproduction methods. 2018, bioRxiv 2020.03.31.011650. bioRxiv. Available online: website page. (accessed on day month year)
16. Ivanic, J.; Ruedenberg, K. Rotation matrices for real spherical harmonies. direct determination by recursion. *J. Phys. Chem.* **1996**, *100*, 6342–6347, doi:10.1021/jp9833350.
17. Pinchon, D.; Hoggan, P.E. Rotation matrices for real spherical harmonics: General rotations of atomic orbitals in space-fixed axes. *J. Phys. A Math. Theor.* **2007**, *40*, 1597–1610, doi:10.1088/1751-8113/40/7/011.
18. Schörkhuber, C.; Zaunschirm, M.; Höldrich, R. Binaural Rendering of Ambisonic Signals via Magnitude Least Squares. *Proc. DAGA* **2018**, *44*, 339–342.
19. Pollow, M.; Klein, J.; Dietrich, P.; Vorlaender, M. *Including Directivity Patterns in Room Acoustical Measurements*; Acoustical Society of America: Melville, NY, USA. 2013; Volume 015008, pp. 015008, doi:10.1121/1.4800303.
20. Noisternig, M.; Klein, J.; Berzborn, M.; Recher, A.; Warusfel, O. High-Resolution MIMO DRIR Measurements in an Opera Hall. In proceedings of the 42nd Annual German Congress on Acoustics (DAGA), Aachen, Germany, 14-17 Mar 2016.
21. Zaunschirm, M.; Baumgartner, C.; Schörkhuber, C.; Frank, M.; Zotter, F. An Efficient Source-and-Receiver-Directional RIR Measurement Method. In *Proceedings of the DAGA*; DAGA, 2017; pp. 1343–1346.
22. Zaunschirm, M.; Frank, M.; Zotter, F. Binaural Rendering with Measured Room Responses: First-Order Ambisonic Microphone vs. Dummy Head. *Appl. Sci.* **2020**, *10*, 1631. doi:10.3390/app10051631.
23. Tervo, S.; Patynen, J.; Lokki, T. Acoustic reflection path tracing using a highly directional loudspeaker. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 18–21 October 2009; pp. 245–248.
24. Zotter, F.; Zaunschirm, M.; Frank, M.; Kronlachner, M. A Beamformer to Play with Wall Reflections: The Icosahedral Loudspeaker. *Comput. Music J.* **2017**, *41*, doi:10.1162/comj_a_00429.
25. Zaunschirm, M.; Frank, M.; Zotter, F. An Interactive Virtual Icosahedral Loudspeaker Array. In *Proceedings of the DAGA*; DAGA: Aachen, Germany, 2016; pp. 1331–1334.
26. Pelzer, S.; Pollow, M.; Vorländer, M. Auralization of a virtual orchestra using directivities of measured symphonic instruments. In Proceedings of the Acoustics 2012 Nantes Conference, Nantes, France, 23–27 April 2012; pp. 2379–2384.
27. Allen, J.B.; Berkley, D.A. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.* **1979**, *65*, 943–950, doi:10.1121/1.382599.
28. Keller, J.B. Geometrical theory of diffraction. *J. Opt. Soc. Am.* **1962**, *52*, 116—-130, doi:10.1007/BF02846778.
29. Svensson, U.P.; Fred, R.I.; Vanderkooy, J. An analytic secondary source model of edge diffraction impulse responses. *J. Acoust. Soc. Am.* **1999**, *106*, 2331–2344, doi:10.1121/1.428071.
30. Tsingos, N.; Funkhouser, T.; Ngan, A.; Carlbom, I. Modeling acoustics in virtual environments using the uniform theory of diffraction. In Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2001, Los Angeles, CA, USA, 12–17 August 2001; pp. 545–553, doi:10.1145/383259.383323.
31. Vorlaender, M. *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*; RWTHedition; Springer: Berlin/Heidelberg, Germany, 2008, doi:10.1007/978-3-540-48830-9.
32. Schoerkhuber, C.; Hoeldrich, R. Signal-Dependent Encoding for First-Order Ambisonic Microphones. In Proceedings of the 43rd DAGA Conference, Kiel, Germany, 6–9 March 2017.
33. Massé, P.; Carpentier, T.; Warusfel, O.; Noisternig, M. Denoising directional room impulse responses with spatially anisotropic late reverberation tails. *Appl. Sci.* **2020**, *10*, 1033. doi:10.3390/app10031033.
34. Tuncer, T.E.; Friedlander, B. *Classical and Modern Direction-of-Arrival Estimation*; Academic Press: Cambridge, MA, USA, 2009.

35. Jarrett, D.P.; Habets, E.A.P.; Naylor, P.A. 3D Source localization in the spherical harmonic domain using a pseudointensity vector. In proceedings of the 18th European Signal Processing Conference, Aalborg, Denmark, 23-27 August, 2010.
36. Zotter, F.; Frank, M. *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*; SpringerOpen: Berlin/Heidelberg, Germany, 2019; pp. 1–210, doi:10.1007/978-3-030-17207-7.
37. Politis, A.; Delikaris-Manias, S.; Pulkki, V. Direction-of-arrival and diffuseness estimation above spatial aliasing for symmetrical directional microphone arrays. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing—Proceedings*; IEEE: Piscataway, NJ, USA, 2015; pp. 6–10, doi:10.1109/ICASSP.2015.7177921.
38. Zaunschirm, M.; Frank, M.; Zotter, F. BRIR Synthesis Using First-Order Microphone Arrays. In Proceedings of the Conference of the Audio Eng. Soc. 144, Milan,Italy, 23–26 May 2018; pp. 1–10.
39. Unsold, A. Beiträge zur quantenmechanik der atome. *Annalen der Physik* **1927**, *387*, 355—-393.
40. Zotter, F. A Linear-Phase Filter-Bank Approach to Process Rigid Spherical Microphone Array Recordings. In Proceedings of the 5th IcETRAN, Subotica, Serbia, 11–14 June 2018; pp. 550–557.
41. Zaunschirm, M.; Schörkhuber, C.; Höldrich, R. Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint. *J. Acoust. Soc. Am.* **2018**, *143*, 3616–3627, doi:10.1121/1.5040489.
42. Zaunschirm, M.; Frank, M.; Franz, Z. Perceptual Evaluation of Variable-Orientation Binaural Room Impulse Response Rendering. In 2019 AES International Conference on Immersive and Interactive Audio, York, UK, 27-29 March, 2019.
43. Merimaa, J.; Pulkki, V. Spatial impulse response rendering I: Analysis and synthesis. *J. Audio Eng. Soc.* **2005**, *53*, 1115–1128.
44. Daniel, J.; Rault, J.; Polack, J. Ambisonics encoding of other audio formats for multiple listening conditions. In *Audio Engineering Society Convention 105*; 1998.
45. Zotter, F.; Frank, M. All-round ambisonic panning and decoding. *J. Audio Eng. Soc.* **2012**, *60*, 807–820.
46. Plenge, G. On the problem of 'in head localization'. *Acta Acust. Acust.* **1972**, *26*, 241–252, doi:10.1007/BF00026991.
47. Werner, S.; Klein, F.; Mayenfels, T.; Brandenburg, K. A summary on acoustic room divergence and its effect on externalization of auditory events. In Proceedings of the 2016 8th International Conference on Quality of Multimedia Experience, QoMEX 2016, Lisbon, Portugal, 6–8 June 2016, doi:10.1109/QoMEX.2016.7498973.
48. Cubick, J. Investigating distance perception, externalization and speech intelligibility in complex acoustic environments Hearing. Ph.D. Thesis, Technical University of Denmark, Lyngby, Denmark, 2017.
49. Wright, M.; Freed, A. Open SoundControl: A New Protocol for Communicating with Sound Synthesizers Matthew. In *International Computer Music Conference (ICMC)*; Michigan Publishing: Thessaloniki, Greece, 1997.
50. Rakerd, B.; Hartmann, W.M. Localization of sound in rooms, III: Onset and duration effects. *J. Acoust. Soc. Am.* **1986**, *80*, 1695–1706, doi:10.1121/1.394282.
51. Anderson, T.W. *An Introduction to Multivariate Statistical Analysis*, 3rd ed.; Wiley-Interscience: New York, NY, USA, 2003; p. 752, doi:10.1080/00401706.1986.10488123.
52. Lindau, A.; Erbes, V.; Lepa, S.; Maempel, H.J.; Brinkman, F.; Weinzierl, S. A spatial audio quality inventory (SAQI). *Acta Acust. Acust.* **2014**, *100*, 984–994, doi:10.3813/AAA.918778.
53. Wilcoxon, F. Individual comparisons by ranking methods. In *Breakthroughs in Statistics*; Springer: Berlin/Heidelberg, Germany, 1992; pp. 196–202.
54. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
55. Meyer-Kahlen, N.; Zotter, F.; Pollack, K. Design and Measurement of First-Order, Horizontally Beam-Controlling Loudspeaker Cubes. In Proceedings of the 144th Convention of the AES, Milan, Italy, 23-26 May, 2018.

2.2 An Interactive Virtual Icosahedral Loudspeaker Array

This work was published as:

M. Zaunschirm, M. Frank, and F. Zotter. An interactive virtual icosahedral loudspeaker array. Fortschritte der Akustik, DAGA, Aachen (2016).

The idea and concept of this article were outlined by me, the first author, with help from the second, and third author. I wrote the original draft of the manuscript with periodical contributions from the third and second author. I did most of the programming and graphical work, and prepared the samples for the listening experiment. I programmed and designed the listening experiment with periodic contributions from the second and third author. The listening experiment was conducted and evaluated by me with periodic contributions from the second author.

An Interactive Virtual Icosahedral Loudspeaker Array

Markus Zaunschirm, Matthias Frank, Franz Zotter
Institute of Electronic Music and Acoustics, University of Music and Performing Arts, 8010 Graz, Austria.
Email: zaunschirm@iem.at

Introduction

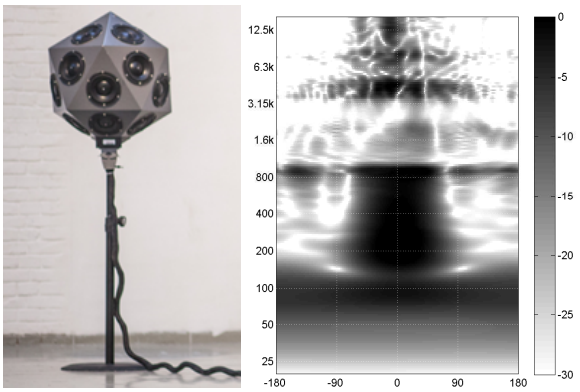
Loudspeaker systems with controllable directivity are new tools to influence the presentation of sound in a room. By individually controlling the strengths of acoustic propagation paths (direct and reflected sound) such loudspeakers allow to create auditory objects. As one example of such a device, our compact icosahedral loudspeaker system (ICO) has been employed in various concerts at different venues. At each venue, the electroacoustic compositions have been adjusted to cause similar auditory objects in the specific acoustic environment. For these artistically important adjustments, despite typically limited, a reasonable amount of time is required in which the venue and the ICO are available.
We discuss the interactive virtualization of the ICO for different playback/recording positions at various venues, to be used on a simple laptop equipped with a spatial audio workstation (Reaper with ambiX plugin suite) and headphones. The virtual ICO (VICO) employs multiple input multiple output (MIMO) convolutions comprising measured room impulse responses (RIRs), impulse responses (IRs) controlling the ICO and the higher-order microphone array (*Eigenmike EM32*), and individual head-related IRs (HRIRs) used for dynamic binaural *Ambisonic* rendering. We finally present an evaluation study that consists of listening experiments comparing the ICO and its virtualized counterpart.

The ICO

The ICO, a compact spherical loudspeaker array, houses 20 loudspeakers that are mounted into the rigid facets of an icosahedron, see fig. 1(a). Spherical beamforming of the ICO allows to synthesize well defined directivities that can be adjusted uniformly to any desired direction defined by azimuth and zenith angle. Starting from controlling the sound particle velocity on the surface of the ICO in terms of spherical harmonics, a sound pressure pattern at any radius can be derived. Accordingly, the pattern undergoes a radius- and frequency-dependent transition as it propagates. The transition of the beam pattern on the surface of the ICO to its far field counterpart is described by superposition of frequency responses for each spherical harmonic order n

$$b_n(k\mathrm{R}) = \frac{\rho c}{\mathrm{i}} \frac{\mathrm{i}^{n+1}}{k\, h_n'^{(2)}(k\mathrm{R})}, \qquad (1)$$

where ρc is the density of air $1.2\,\mathrm{kg/m^3}$ times speed of sound $343\,\mathrm{m/s}$, $\mathrm{i} = \sqrt{-1}$ is the imaginary unit, $k = 2\pi f/c$ is the wave number with the frequency f, and $h_n'^{(2)}(k\mathrm{R})$ is the derivative of the n^{th}-order spherical *Hankel* function of the second kind, with $\mathrm{R} = 28.5\,\mathrm{cm}$ as acoustically effective radius of the ICO. The patters are strongly attenuated as the order n increases and the frequency decreases. In order to compensate for the occurrent magnitude and phase changes an equalization by *far-field radiation control* is needed, cf. fig. 2.
Moreover, the loudspeakers of the ICO share a common enclosure and thus, their motions are acoustically coupled. As directivity pattern synthesis requires individual control of these motions, a *MIMO cross-talk canceller* is required in order to compose far-field patterns out of superimposed spherical harmonics.
With the suitable control system (*far-field radiation control* and *MIMO cross-talk canceller*) the creation of desired directivities in terms of spherical harmonics is achieved by using the same tools as for arranging sounds in higher-



(a) Staging of the ICO. (b) Far-field beam pattern.

Figure 1: Staging of the ICO (left) and horizontal cut through far-field beampattern with magnitude in dB (grayscale) over polar angle and frequency for spherical beamforming with the ICO using radiation control and acoustic cross-talk cancellation (right).

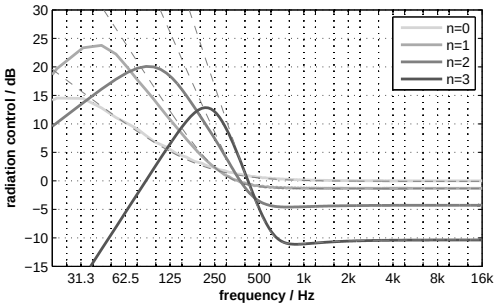


Figure 2: Analytic (dotted) and practical (solid) radial filters for the ICO. Cut-off frequencies for limiting the higher order signal amplitudes are set to [20,37,115,240] Hz.

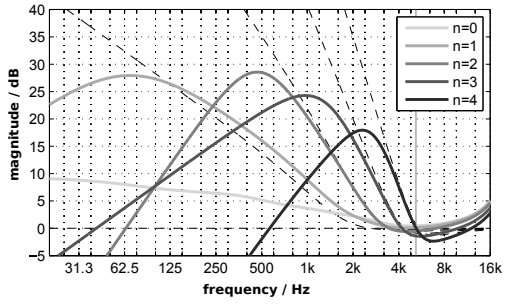


Figure 4: Analytic (dotted) and practical (solid) holographic filters for a spherical rigid sphere microphone array up to spherical harmonics order $N = 4$. Lowest possible cut-off frequencies that limit the WNG to a maximum of 20dB are $[20, 59, 612, 1534, 2742]$Hz.

order *Ambisonic*, i.e. *Ambisonic encoding*.
Fig. 1(b) shows the far-field beam pattern steered towards an angle of 0° on the horizon that is obtained using the above described processing. At lower frequencies the order of the directivity pattern is limited as inversion of b_n for $n > 1$ would require such a strong bass boost that would damage the loudspeakers in practice. For frequencies above 150 Hz we get a 3^{rd}-order beam pattern until the spatial aliasing becomes dominant above 800 Hz.

The virtual ICO

The signal processing chain of the interactive virtual ICO is depicted in fig. 3. A spatial arrangement of sound objects x_S, according to the desired directions φ_S, ϑ_S, is achieved by 3^{rd}-order *Ambisonic encoding*. The driving signals of the ICO's loudspeakers are then obtained by a frequency-independent decoder and an equalization using radial and crosstalk cancellation filters [1]. Until this point the processing for the ICO and its virtualized version is identical. For virtualization, the ICO to EM32 block contains a dataset of virtualized/measured MIMO room impulse responses (RIRs) of a performance situation (with an EM32 positioned at a desired listening position). Additional information on measurement configuration and data can be found here [2].

Holographic Filters
To allow for *Ambisonic* surround playback, the 32-channels of the Eigenmike EM32 are first transformed into the spherical-harmonics domain by a frequency-independent *encoder* matrix. However, the obtained signals are not suitable for direct playback as the low frequency components of the signals are mainly mapped to a omni-directional pattern (similar to the attenuation of higher-order components for the ICO) on the surface of the microphone. To compensate for the attenuation holographic filters that yield the desired *Ambisonic* playback signals are needed [3]. The frequency responses of the employed holographic filters are depicted in fig. 4, for further reading see [4].
The decomposed *Ambisonic* signals are rotated dynamically according to head movements of the listener and are finally decoded using customizable HRIRs before playback via headphones (*Ambsionic* signals of Eigenmike can also be decoded to any loudspeaker setup using [5]).

Listening experiments

Previous work shows how sound objects generated by the ICO are perceived by listeners [6]. In this experiment we evaluated the differences of the sound objects created by the real ICO and its virtualized counterpart. The experiments are conducted in a real room with dimensions of 6.8 m×7.6 m×3 m and a mean reverberation time of 0.57 sec. (IEM lecture room). The ICO was placed near the front right corner of the room and the listening position is chosen approximately 4 m away form the ICO, see fig. 7.

Room Database
The 20×32 RIRs between the ICO and the Eigenmike are measured using the exponentially-swept sine technique [7] with a 5 sec. long sweep. Despite using relatively long sweeps we experienced unrealistically long reverberation times in the measured RIRs, especially for frequencies > 4 kHz. This can be explained by the limited efficiency of the ICO transducers in that frequency range. Thus, time-frequency filtering (de-noising) of the measured RIRs was needed.
Let us assume that the short-time fourier transform (STFT) of the late part of a RIR is well modeled by

$$H(t,\omega) = uv^{-t}N_1(t,\omega) + wN_2(t,\omega), \tag{2}$$

where $N_{1,2}(t,\omega)$ are two uncorrelated normalized noise processes, $v > 1$, and (t,ω) represents time and frequency dependency of the STFT, respectively. The expected value of the summed squared amplitudes (for all 640 paths) is derived from eq. 2 as

$$E\{|H(t,\omega)|^2\} = u^2v^{-2t} + w^2, \tag{3}$$

and the time where both processes are equally loud is found by

$$t_0 = \frac{\ln \frac{w}{u}}{\ln v}. \tag{4}$$

Unnatural and thus unwanted background noise present in the RIRs is adapted to the slope of reverberation by using

$$H_{denoise}(t,\omega) = \frac{H(t,\omega)}{\sqrt{1 + v^{2(t-t_0)}}}, \tag{5}$$

where the model parameters u, v, w are found at each frequency ω from comparing the modeled energy decay relief (EDR) [8] with that of the measured RIR

$$EDR_{RIR}(t) = \int_T^t |H(t,\omega)|^2 dt, \tag{6}$$

$$EDR_{model}(t) = \frac{u^2v^{-2t}}{-2\ln v} + w^2(T-t). \tag{7}$$

The model parameters defined in eq. 2 can be found in suitable sections of EDR_{RIR}, see vertical black lines in fig. 5. Typically, the parameters u and v can be estimated by linear regression of the early part of $10 \lg EDR_{RIR}$ with

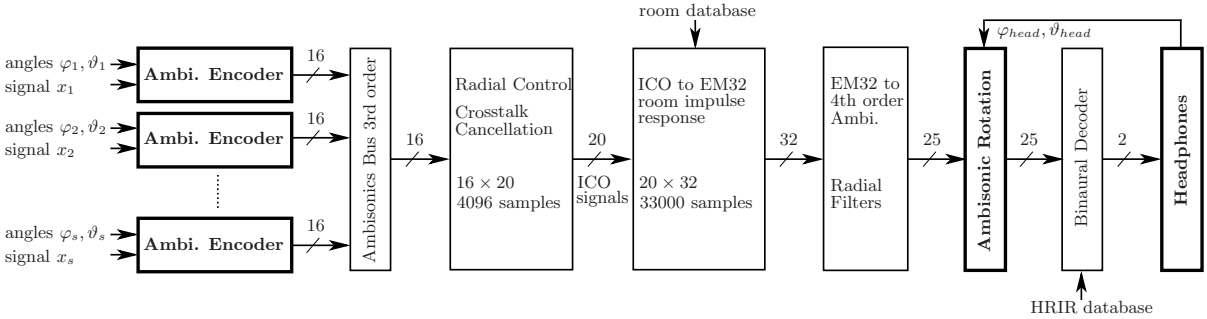


Figure 3: Processing chain of the virtual ICO containing real-time encoding and head-tracked headphone playback for user interaction (bold blocks). Sounds are spatially arranged according to the desired directions using the *Ambi. Encoder*.

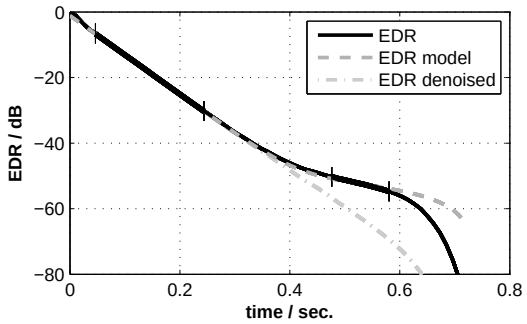


Figure 5: *EDR* of the actual, modeled and de-noised RIR for a third-octave band with center frequency $\omega = 1024\,\mathrm{Hz}$. The section for model parameter estimation is marked by vertical black lines.

$10 \lg EDR_{model} = -20t \lg v + 10 \lg \frac{u^2}{-2 \ln v}$. The parameter w can be found by regression of the later part of EDR_{RIR} by assuming $EDR_{model} = w^2(T - t)$ in that part. Figure 5 shows the EDR of the original, modeled and de-noised RIR for a third-octave band with a center frequency $\omega = 1024\,\mathrm{Hz}$.

Stimuli and Experiment Setup

The test signals are 1.5 sec. long and include a pink noise burst with attack and release times of $t_a = t_r = 500\,\mathrm{ms}$ and a sequence of irregular short bursts. Both sounds are steered on the horizon towards 0°, 70°, 180° and −90° using a 3^{rd} order directivity pattern. The play-back signals are rendered in real time using *Reaper*, the *ambiX* and *mcfx* plugins [9] for both, the ICO and the VICO system, cf. fig. 3. Listeners are asked to specify the location of the perceived auditory object for each of the 8 (2 sounds and 4 beam directions) conditions by placing a marker corresponding to the played back sound using a GUI that includes a map of the listening room. Listeners could switch between all 8 conditions on one page arbitrarily and listen to or compare them as often as wanted. Presentation methods included playback with the real ICO and the VICO using head-tracked binaural rendering with *AKG K712* headphones. Independent of the playback method the listeners were seated at the same position and thus, could also see the real ICO for presentation using the VICO environment. For head tracking we used a *OptiTrack* system consisting of 6 *Flex 13* cameras[1]. The

[1]A suggestion of a smaller head tracking device can be found at `http://www.matthiaskronlachner.com/`

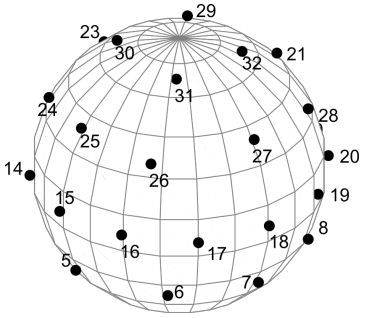


Figure 6: Virtual loudspeaker setup used for binaural rendering. The setup consists of four rings at a zenith angle of $[120°, 90°, 60°, 25°]$ including $[8, 12, 8, 4]$ loudspeakers with a spacing of $[45°, 30°, 45°, 90°]$.

Ambisonic signals were rotated contrary to head movements of listeners using an update rate of 100 Hz.

The participants were grouped such that 4 of the overall 8 participants rated first the ICO and then the VICO. The other group rated in vice versa playback order. Each playback method was repeated three times with randomly ordered conditions in each trial (3 times the 8 conditions played back with the ICO followed by 3 times the 8 conditions played back with the VICO). In each of the two groups we had two participants using individual HRIRs for playback of the VICO.

Individual in-the-ear HRIRs were measured with the blocked-ear-canal technique using a *Sennheiser KE-4-211-2* microphone and with the multiple exponential sweep method [10] in a semi-anechoic chamber. Overall, the measurement grid consisted of 1550 positions with a resolution of 2.5° and 10° in azimuth and zenith, respectively (loudspeakers were arranged on a sphere with a radius of 1.2 m). Out of the entire set of HRIRs we chose 32 that corresponded to the locations of the 32 virtual loudspeakers that are used for *Ambisonic* decoding, see fig.6. The actual frequency-independent decoder matrix is calculated according to the *All-Round Ambisonic Decoding* (AllRAD) strategy [11].

In order to evaluate not only the location of the perceived auditory object but also the naturalness of the acoustic scene for playback within the VICO environment we conducted a plausibility test. The irregular noise burst signal was steered towards the listening position and 4 partici-

pants were asked to state if they listen to the real ICO or the VICO. This time participants had to wear the headphones during the entire test. In each trial we presented the stimulus ten times in random order: five times with the ICO and VICO, respectively. The playback signal of the real ICO was adapted such that high-frequency damping of the headphone shell is accounted for.

Results

Generally, participants reported a highly natural sounding acoustic scene for playback within the VICO environment. These comments are underlined by the result of the plausibility test: only 52.5% of the ratings were correct.
Figure 7 shows the 95% confidence area of the mean localization for each of the 8 conditions. It can be seen that both direction and distance are well perceivable with the virtualized version of the ICO. In detail, ratings of the perceived distance were not significantly different among all conditions. The perceived direction was only significantly different for two conditions, see bold entries in tab. 1. For both playback methods irregular noise bursts tend to be localized nearer to the ICO/VICO than the stationary noise signal. This can be explained by a more prominent precedence effect for highly transient signals than for stationary signals [12].
Neither the use of individual HRIRs nor whether ICO or VICO was presented first had any significant impact on the ratings.

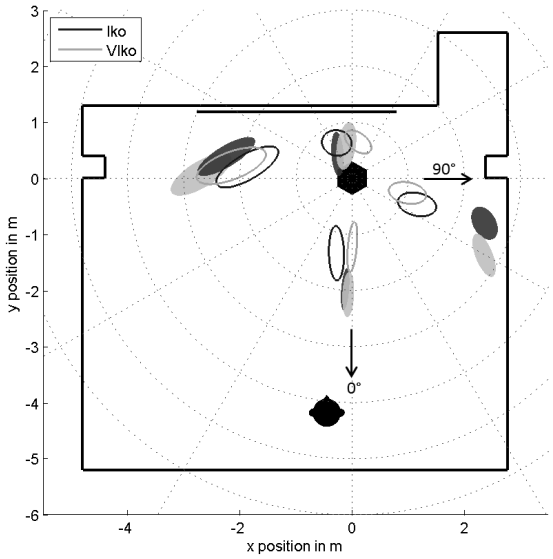


Figure 7: Setup and results of the listening experiment. The ICO (black filled hexagon) is placed in the front right corner of the room. The listening position is indicated by filled head symbol. 95% confidence ellipsoids of subjects ratings are shown for ICO (dark gray) and VICO (light gray) for a noise (filled) and irregular burst (not filled) signal.

Table 1: p-values for differences in direction and distance ratings obtained for listening experiments comparing ICO and VICO as playback device. Conditions $1-4$ and $5-8$ refer to static noise signal and irregular noise bursts, respectively. Within the signal groups the direction of the beam is altered clockwise starting from $-90°$ (left).

	condition							
	1	2	3	4	5	6	7	8
direction	0.28	0.35	**0.05**	0.70	0.71	**0.05**	0.36	0.13
distance	0.24	0.24	0.22	0.71	0.62	0.81	0.32	0.37

Conclusion

We presented an interactive virtualization of the ICO. Listening experiments showed that the perceived auditory objects are comparable to those obtained by the ICO. Thus, the VICO allows to evaluate new MIMO filters (for ICO and Eigenmike) and it can be used to more easily master or preproduce musical pieces for different rooms.

Acknowledgments

This work is part of the project Orchestrating Space by Icosahedral Loudspeaker (OSIL), which is funded by the Austrian Science Fund (FWF): PEEK AR 328. We thank all listeners for their participation in the experiments.

References

[1] S. Lösler: *MIMO-Rekursivfilter für Kugelarrays*. Master thesis, 2014.

[2] Orchestrating Space by Icosahedral Loudspeaker, URL: `http://iem.kug.ac.at/osil/`

[3] J. Daniel and S. Moreau: Further Study of Sound Field Coding with Higher Order Ambisonics. Proc. of the 116th Convention of the Audio Eng. Soc., pp. 1–14, 2004.

[4] S. Lösler and F. Zotter: Comprehensive Radial Filter Design for Practical higher-order Ambisonic Recording. Fortschritte der Akustik, DAGA, pp. 452–455, 2015.

[5] F. Zotter and M. Frank: All-round ambisonic panning and decoding. AES: Journal of the Audio Engineering Society, vol. 60, no. 10, pp. 807–820, 2012.

[6] M. Frank, G. K. Sharma, and F. Zotter: What we already know about spatialization with compact spherical arrays as variable-directivity loudspeakers. Proc. inSONIC2015, 2015.

[7] A. Farina: Simultaneous measurement of impulse response and distortion with a swept-sine technique. Proc. AES 108th conv, no. I, pp. 1–15, 2000.

[8] J. M. Jot: An analysis/synthesis approach to real-time artificial reverberation. Proceedings ICASSP-92, pp. 221–224 vol.2, 1992.

[9] M. Kronlachner: Plug-in Suite for Mastering the Production and Playback in Surround Sound and Ambisonics. 136th AES Convention, April 2014, pp. 3–7, 2014.

[10] P. Majdak, P. Balazs, and B. Laback: Multiple exponential sweep method for fast measurement of head-related transfer functions. Journal of the Audio Engineering Society, vol. 55, no. 7-8, pp. 623–636, 2007.

[11] F. Zotter: Holofonie für Musikinstrumente. Fortschritte der Akustik - DAGA, 2012.

[12] B. Rakerd and W. M. Hartmann: Localization of sound in rooms, II: The effects of a single reflecting surface. The Journal of the Acoustical Society of America, vol. 78, pp. 524–533, 1985.

3

Beamforming with the Icosahedral Loudspeaker Array

3.1 A Beamformer to Play with Wall Reflections: The Icosahedral Loudspeaker

This work was published as:

F. Zotter, **M. Zaunschirm**, M. Frank, and M. Kronlachner. A Beamformer to Play with Wall Reflections: The Icosahedral Loudspeaker. *Computer Music Journal*, 41(3), 2017. ISSN 15315169. doi:10.1162/comj_a_00429.

The idea and concept of this article were outlined by the first author. I, as second author, was involved in the writing of the original draft and contributed significantly to the virtual IKO for headphones and to the measurements, programming, and formalism presented in the second part of the article: **Beamformer Theory and Control of the IKO**. The first author and I were writing the revised version with periodic contributions of the third author.

A Beamformer to Play with Wall Reflections: The Icosahedral Loudspeaker

Franz Zotter,* Markus Zaunschirm,* Matthias Frank,* and Matthias Kronlachner†

*Institute of Electronic Music and Acoustics
University of Music and Performing Arts
Inffeldgasse 10/3
8010 Graz, Austria
{zotter, zaunschirm, frank}@iem.at
†Automotive Systems GmbH
Harman Becker
Schlesische Str. 135
94315 Straubing, Germany
matthias.kronlachner@harman.com

Abstract: The quote from Pierre Boulez, given as an epigraph to this article, inspired French researchers to start developing technology for spherical loudspeaker arrays in the 1990s. The hope was to retain the naturalness of sound sources. Now, a few decades later, one might be able to show that even more can be done: In electroacoustic music, using the icosahedral loudspeaker array called IKO seems to enable spatial gestures that enrich alien sounds with a tangible acoustic naturalness.

After a brief discussion of directivity-based composition in computer music, the first part of the article describes the technical background of the IKO, its usage in a digital audio workstation, and psychoacoustic evidence regarding the auditory objects the IKO produces. The second part deals with acoustic equations of spherical beamforming, how the IKO's loudspeakers are controlled correspondingly, how we deal with excursion limits, and the resulting beam patterns generated by the IKO.

> The loudspeaker "anonymizes" the actual source. . . . There will be more resemblance, in a certain way, between amplified piano and amplified harp, than between amplified and unamplified piano. One could say that the instruments have gone through a "rolling mill" of amplification and have lost some of their individuality. . . . The composer is left to play with this phenomenon and to make use of it in an informed manner.
>
> —Boulez 1983

Composing Directivity for Electroacoustic Music

In electroacoustic music, we find the application of directionality in Marco Stroppa's music that used vertically stacked loudspeakers, each of which was aimed at a different angle. These were called *totem acoustiques*, and were used in his compositions *. . . of Silence* (2007), *Hist Whist* (2009), and most prominently in the opera *Re Orso* (2011) with an eight-loudspeaker column hanging in the middle of the stage.

A starting point to composing with directivity is "La Timée," a cube housing six loudspeakers utilized by the researchers at IRCAM in order to give more naturalness to loudspeaker-based diffusion of sounds (Caussé, Bresciani, and Warusfel 1992; Warusfel, Derogis, and Caussé 1997; Misdariis et al. 2001).

The playback system discussed here, called IKO, is a 20-sided, 20-channel loudspeaker system in the form of the convex regular icosahedron (see Figure 1). As a compact spherical loudspeaker array, the IKO provides the technical means to project a focused sound beam in a freely adjustable direction. Inside a room, this kind of beam direction can be set to predominantly excite selected wall reflections, or combinations of reflections, causing interesting effects in perceived localization.

Although the beamforming of the IKO is capable of uniform adjustment to all directions, contiguous directions are not mapped to contiguous perceived directions, as reflection paths of a room are discrete. Still, the IKO's sculptural auditory objects

Computer Music Journal, 41:3, pp. 50–68, Fall 2017
doi:10.1162/COMJ_a_00429

Figure 1. The IKO is a 20-sided, 20-channel loudspeaker array in the form of an icosahedron. Its diameter is about 60 cm.

(cf. Sharma 2016) offer an exciting spatialization technology to composers of electroacoustic music, in the broadest sense of the term. The IKO's presence on stage offers a scene that is unexpectedly pronounced and natural, like that of a human performer.

We promote the use of variable directivity of compact spherical loudspeaker arrays in computer music to create new auditory objects. The first half of this article is dedicated to the existing software solutions to working with the IKO, the IKO's hardware, its staging, and what is known about the perception of its sound beams in a room. The second half provides a deeper understanding of the acoustic principles behind the IKO's spherical beamforming, an approach to stay within excursion limits of the transducers, and details on how the spherical beamforming and velocity control of the IKO are achieved and verified, on the basis of measurements and a multiple-input, multiple-output (MIMO) system design.

Part I: Beamforming and the IKO in Practice

The first half of this article deals with a literature review on beamforming, beamforming applications, and compact spherical loudspeaker arrays, and information about the IKO that is relevant for its practical application—how it is controlled with plug-ins for a digital audio workstation (DAW), how it is virtualized for different rooms using binaural synthesis, what it is built of, how it can be presented and staged, and which auditory objects can be expected from the existing perceptual studies.

Technical Background and Literature

Classical beamforming technology aims at focused emission and reception of waves by arrays of transducers driven and superimposed with different weights, delays, or filters. If beamforming only uses delays or weights, we speak of *delay-and-sum* or *weight-and-sum* beamforming, respectively, whereas the most general approach using filters is called *filter-and-sum* beamforming (Schelkunoff 1943; Brandstein and Ward 2001). When allowing filters with gains exceeding the maximum of the directivity pattern, strong focusing is possible even with small apertures. This is called *superdirective* or *supergain* beamforming (Bloch, Medhurst, and Pool 1953; Elko 2000).

Based on the idea of exploiting beamforming to selectively excite wall reflections as a type of surround-sound technology, a planar loudspeaker array at the typical center loudspeaker position is commercialized in Yamaha's Sound Bar for home-cinema applications (Takumai 2006). In this application, surround and side loudspeakers are substituted by beams emphasizing suitable wall reflections.

As an alternative method of creating strongly focused sound beams, parametric arrays utilize the nonlinearity of air (Bennett and Blackstock 1975; Croft and Norris 2003). A powerful group of ultrasound transducers plays an amplitude-modulated carrier frequency above 35 kHz at a sound pressure level higher than 120 dB. Its envelope is demodulated along the propagation path. The interpretation as a nonlinear source phenomenon is called a *parametric array*. Sugibayashi et al. (2012) built and evaluated the use of directionally adjustable ultrasound transducer arrays mounted on each of the 20 surfaces of an icosahedron to establish a mixed-reality sound-field synthesis. This intriguing system had to be supplemented, however, by common electrodynamic transducers to support frequencies below 1 kHz.

The technology utilized for compact spherical loudspeaker arrays such as the IKO is linear and superdirective, and is called *spherical harmonic beamforming* (cf. Butler and Ehrlich 1977 for underwater sound and Warusfel, Derogis, and Caussé 1997 for music). The targeted directional resolution is uniform and independent of the beam direction, and the technique is based on filtering to equalize different attenuations for spherical harmonics of different orders, when radiated to the near or far field (Zotter and Noisternig 2007).

The reasoning behind a variable-directivity playback device for music (as presented by Caussé, Bresciani, and Warusfel 1992; Warusfel, Derogis, and Caussé 1997; Misdariis et al. 2001) has motivated other research groups to pursue technical efforts of establishing and controlling compact spherical loudspeaker arrays. At Princeton University, arrays like these have been built for electroacoustic performances with, for instance, the Princeton laptop orchestra (Cook et al. 1998; Trueman et al. 2006). At the University of California, Berkeley, researchers investigated magnitude-based beam-pattern control and accuracy limits (Kassakian and Wessel 2003, 2004; Kassakian 2005, 2006; Avizienis et al. 2006). Further notable efforts to build, control, and use arrays such as these have been undertaken in Austria (Zotter and Höldrich 2007; Zotter and Noisternig 2007; Pomberger 2008; Zotter, Pomberger, and Schmeder 2008; Zotter 2009; Kerscher 2010; Zotter and Bank 2012), Germany (Pollow and Behler 2009; Pollow 2014), Brazil and France (Pasqual 2010; Pasqual, Herzog, and Arruda 2010; Pasqual, Arruda, and Herzog 2010), Israel (Rafaely and Kaykin 2011; Morgenstern, Zotter, and Rafaely 2012; Morgenstern, Rafaely, and Zotter 2015), Australia (Miranda, Cabrera, and Stewart 2013), and New Zealand (Poletti, Betlehem, and Abhayapala 2015).

Other work has been pursued recently in our present research project Orchestrating Space by Icosahedral Loudspeaker (OSIL), for which the goal is to artistically find and define sound sculptures by composing music with the IKO (Sharma 2016) and to scientifically investigate these sculptures (Sharma, Zotter, and Frank 2014; Frank, Sharma, and Zotter 2015; Wendt et al. 2016, 2017; Zaunschirm, Frank, and Zotter 2016).

Controlling the IKO with DAW Plug-ins

Real-time performances and composition for the IKO can be controlled from a standard consumer personal computer, or even a laptop. The 20 driving signals for the IKO's loudspeakers are generated using a combination of the ambiX and mcfx VST plug-ins (Kronlachner 2014), see Figure 2. The ambiX software allows one to create, modify, and decode higher-order Ambisonics on a DAW. Although Ambisonics is usually associated with loudspeakers surrounding the audience, the same representation is deployed to control directional beams radiated outwards from compact spherical arrays such as the IKO.

Care has been taken to make the required matrix convolution efficient, as the IKO's control system (see Figure 3) consists of 320 FIR filters (20 channels, each requiring 16 filters) whose coefficients are stored as WAV files. The mcfx_convolver software carries out convolutions as nonuniform, partitioned fast convolutions for low latency and low CPU load. The filters used by mcfx_convolver to connect each of its inputs to each of its output is specified in a configuration file. Selecting from different configuration files allows performance with different setups—for instance, on a different array or with a different filter set. To host the ambiX and mcfx VST plug-ins, DAWs and other music software environments, such as Max, AudioMulch, Bidule, Ardour, and Reaper, must be capable of dealing with at least 20 channels. For the IKO, we mainly use Reaper because of its support for up to 64 channels per track or bus. Furthermore, it supports live performance and improvisation by remote control using Open Sound Control (OSC). Other convenient features include the ability to record and program automations, as well as faster-than-real-time rendering of finished projects to 20-channel audio files. Reaper and the plug-ins are available under Windows, Mac OS, and Wine (Linux).

The signal routing and processing schema is shown in Figure 3. A playback signal $s_c(t)$ and

Figure 2. Screenshots of VST plug-ins controlling the IKO. The plug-in ambix_encoder_o3 (a) controls the beam direction of the IKO for one input signal by encoding it to 16 third-order Ambisonic signals. The plug-in mcfx_convolver24 (b) generates 20 loudspeaker signals from the 16 third-order Ambisonic signals using real-time convolution.

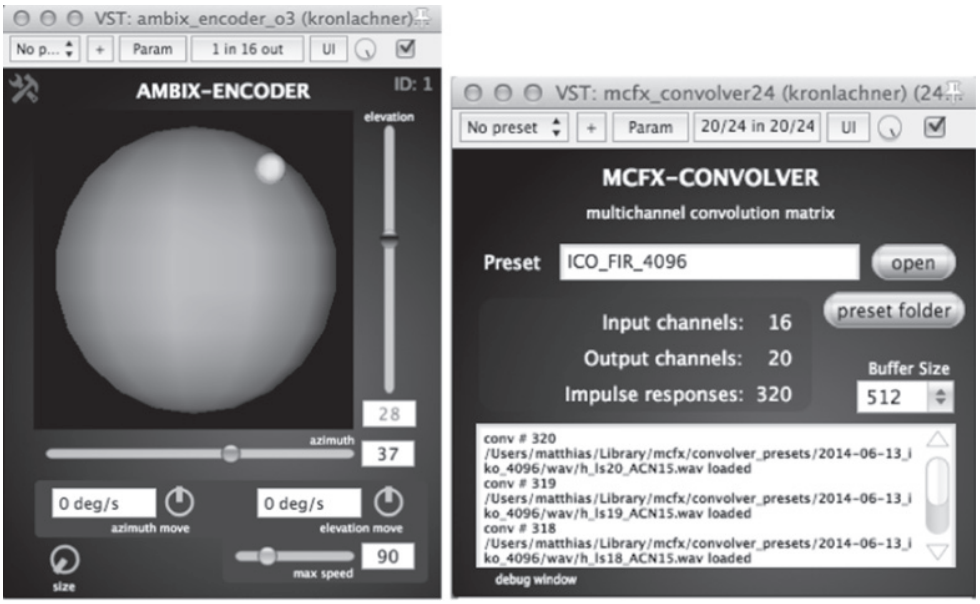


Figure 2

Figure 3. Processing schema controlling the IKO's beamforming. This includes the Ambisonic bus as the sum of encoded source signals, feeding the final MIMO control system.

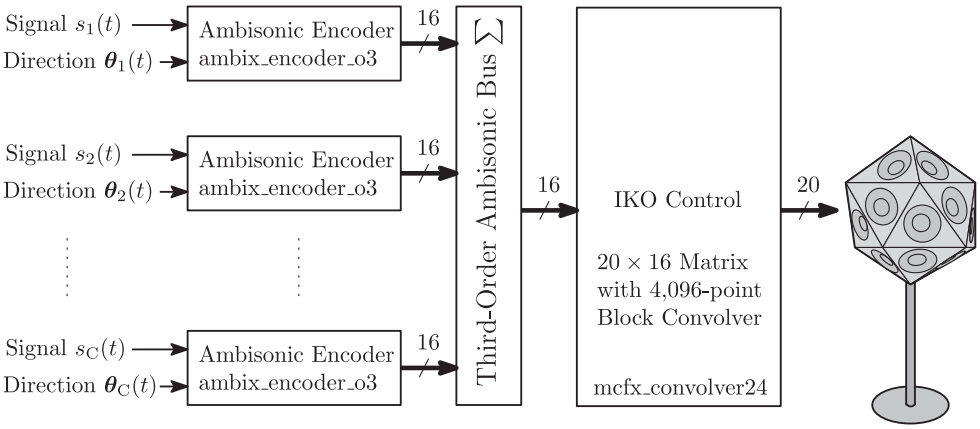
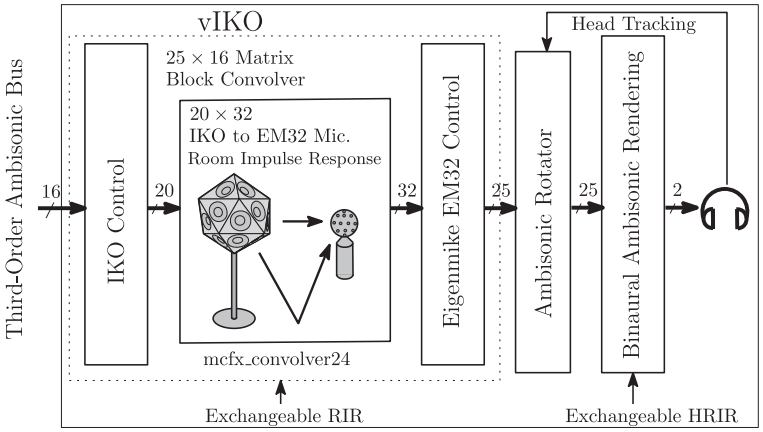


Figure 3

its beam direction $\boldsymbol{\theta}_c(t)$ are fed into a third-order Ambisonics encoder as an insert effect (ambix_encoder_o3). The resulting 16 channels are sent to a master mix. The 20-channel master mix uses a 20×16 fast convolution matrix as an insert effect (mcfx_convolver24), with 4,096 coefficients at a sample rate of 44.1 kHz. The resulting 20 signals feed the amplifiers for the 20 loudspeakers of the IKO. In the real-time operation of the IKO, CPU load amounts to 65 percent for ten sources and a 512-sample buffer, using a MacBook Pro 2.53 GHz Intel Core 2 Duo.

Figure 4. Processing schema of the virtual IKO using measured room impulse responses from the IKO to the Eigenmike EM32 and measured HRIRs.



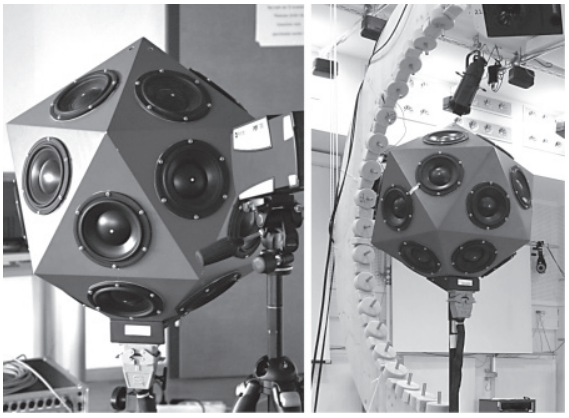
The Virtual IKO for Headphones

The virtual IKO (vIKO) by Zaunschirm, Frank, and Zotter (2016) provides a DAW-based real-time simulation of the IKO by binaural synthesis to headphones, optionally head-tracked (see Figure 4). It provides a Reaper session with suitable routing and delivers a collection of presets for the aforementioned plug-in suites. The presets provided are based on measurements taken in different rooms using the IKO as a source and using the Eigenmike EM32 to capture impulse responses at different listening positions. The vIKO comes with two exemplary room responses and, currently, two sets of head-related impulse responses (HRIRs) measured by the Acoustics Research Institute in Vienna. Each of these sets can be chosen for matrix convolution in the mcfx_convolver plug-in and for binaural rendering in the ambix_binaural plug-in, respectively. Based on vIKO, the OSIL Web site (http://iem.at/osil) offers binaural renderings of basic time-variant beam constructions (called IKO moves) and of musical pieces.

IKO Hardware

The IKO turned out to perform well in electroacoustic concerts thanks to its large and powerful transducers. By contrast, the transducers described in the initial technical report by Zotter and Sontacchi (2007), or smaller arrays such as the prototype by Kerscher (2010), were not powerful enough for concert performances. The IKO is constructed of 20 equilateral triangular faces with 34.7-cm-long outer edges, made from medium density fiberboard, and cut with a bevel of 20.9°. Ten of these faces are glued together in the shape of two pentagonal pyramids of five triangular faces each. The apexes of the two pyramids form the upper and lower apex of the IKO. They are twisted 36° with respect to each other, and they are glued to a horizontal belt of ten faces pointing upward and downward in alternation (cf. Figure 5a). The interior of the IKO is a

Figure 5. Measurement setups for loudspeaker cone velocity and sound pressure: laser vibrometer measurement of voltage-to-loudspeaker cone-velocity transfer functions (a) and sound pressure measurement with semicircular microphone array (r = 75 cm) at IEM CUBE, with a turntable so that directivity is measured at 18 × 36 = 648 directions (b).



(a) (b)

Figure 5

Table 1. Azimuth and Zenith Angles for IKO Loudspeakers

Loudspeaker	*Azimuth*	*Zenith*	*Loudspeaker*	*Azimuth*	*Zenith*
1	0°	142.62°	11	36°	79.19°
2	72°	142.62°	12	108°	79.19°
3	144°	142.62°	13	180°	79.19°
4	−144°	142.62°	14	−108°	79.19°
5	−72°	142.62°	15	−36°	79.19°
6	0°	100.81°	16	36°	37.38°
7	72°	100.81°	17	108°	37.38°
8	144°	100.81°	18	180°	37.38°
9	−144°	100.81°	19	−108°	37.38°
10	−72°	100.81°	20	−36°	37.38°

Loudspeaker 1 is next to the cable socket, the sequence runs counterclockwise from bottom to top.

single unpartitioned volume filled with wool, and it contains the cabling of the transducers, which is attached to the back of a 42-pin Harting Han DD industrial socket. In the center of each of the 20 faces, a 6.3-in. Morel CAW-638 transducer is mounted and can produce an excursion of up to $x_{max} = \pm 4.25$ mm. The channel indices and angular coordinates of the loudspeakers are given in Table 1. The Harting socket on the outside allows one to attach a cable hanging downwards, which is a 15-m-long bundle of 20×1.5-mm^2 loudspeaker wire pairs gathered in a braided sleeving. At the other end, the cables are attached to a socket connected to the 40 banana-jack sockets of a customized sonible d:24. This is a compact 24-channel amplifier consisting of three rack units with 250 W per channel.

The prototype of the IKO, developed at the Institute of Electronic Music and Acoustics (IEM) and described in this article, was used for measurements, experiments, performances, etc., leading to a cooperation with the company sonible to manufacture the IKO and market it commercially (cf. http://iko.sonible.com). This new IKO by IEM and sonible is redesigned for easier transport and for easy integration with the MADI/Dante-capable sonible d:24 multichannel amplifier. It uses a newer Morel transducer series ensuring high performance.

Staging the IKO

The IKO can create auditory objects of high spatial definition when utilizing first-order reflections of the walls in the performance space. In current performance practice, two basic staging constellations are used: one for typical rectangular rooms, and another that uses a concave setup of reflectors behind the IKO (Sharma 2016; see also Figure 6).

Rectangular rooms are the simplest constellation in which the IKO is played, preferably between a corner of the room and the audience. This was the constellation used, for instance, in concerts showcasing IKO held at the International Conference on Digital Audio Effects (DAFx) in 2010, at the Darmstadt International Summer Courses for New Music in 2014, and in the Media Art Gallery at the Zagreb Showroom of Contemporary Sound (Izlog Suvremenog Zvuka) festival in 2015. This arrangement makes it possible to exploit balances between at least two pronounced reflections from the walls and the direct sound. The IKO's distance to the audience should be at least as far as the distance to both walls. Rectangular rooms often offer more effects, such as usable reflections, for example, at the ceiling, at the floor, or from a more distant side wall, as well as spatial reverberation effects. There is, however, some risk depending on the geometry, wall material, sound material, etc. It is therefore

(a) (b)

Figure 6. IKO in the performance setup in MUMUTH, Graz, at the 2010 International Conference on Digital Audio Effects (a), and in the ZKM Kubus, Karlsruhe, at the 2015 InSonic festival (b).

wise to compose pieces that can be adjusted to the given environment.

Alternatively, concave arrangements of reflectors behind the IKO were used at a concert in the Signale Graz festival in 2014 and at the InSonic conference held at ZKM in 2015 (see Figure 6), offering a large set of useful, distinct reflections. We had a similar performance situation in the French Pavilion at the Showroom of Contemporary Sound festival in 2015, where the cylindrical wall of the performance space could be used without modification. The concave reflector arrangement behind the IKO should preferably exhibit a radius of about 5 m to 7 m, and the loudspeaker should be placed in the symmetry axis of the arrangement at a distance of about 1.5 m to 3 m. The concave setup increases the number of reflections (stage wall, side walls), which are otherwise limited, to a plethora of distinct reflections, available everywhere between the stage and the side walls. The audience should preferably be at least about 5 m away from the IKO to allow a balanced perspective on auditory objects that can be shaped by the reflections.

At the low-frequency end (less than 100 Hz), the IKO is omnidirectional and acts as a powerful subwoofer that is well able to excite large spaces. In the octave above 100 Hz, beams radiated by the IKO become directional, so that bass in the octave above 100 Hz can be moved around in the room. Such sounds are often localized as rotary spacious zones that are not colocated with the IKO.

Perception of Sound Beams in Rooms

Although it might seem logical that the sound propagation path emphasized the most would appear as a localized direction in our perception, the precedence effect counteracts this intuition. To study the perceived localization of directional sources with variable orientation, our initial studies (Zotter et al. 2014; Zotter and Frank 2015) considered a simulated source with a third-order beam pattern in a rectangular room. Third-order beam patterns are composed of all spherical harmonics of the orders $n = 0 \ldots 3$. Direct and reflected sound were simulated using the image-source method up to first and second order, which were auralized on 24 Genelec 8020 loudspeakers arranged on a horizontal ring in an anechoic environment.

The first of these two studies showed that the orientation of directional sound sources can be perceived, with a localization that can substantially deviate from the direct path. It used nine volunteer listeners who undertook the task of localizing test signals consisting of bursts of pink noise. The localized direction could be modeled (1) by an extended energy-vector model considering a rough echo threshold of −0.25 dB/ms (Rakerd, Hartmann, and Hsu 2000) and (2) by a binaural predictor based on a model proposed by Werner Lindemann (1986).

The second study, with eleven experienced volunteer listeners with normal hearing, was also based on an auralized source with third-order beam patterns

and tested the perceived direction localization for 36 source orientations in 10° steps. Each listener was presented with conditions in individual random sequence and could respond not only by a single, primary localization direction, but also by a possible secondary one. Listeners were asked to respond by naming integers, based on the even-numbered ticks visibly attached to the loudspeakers. The primary direction could be modeled by the aforementioned extended energy-vector. Secondary directions appeared to be difficult to model and were perceived in only 24 percent of the 36 source orientations using auralization with direct sound and first-order image-sources, but in 42 percent with second-order image sources added, indicating a dependency on later reflections. By contrast, primary localization directions were not changed much by second-order image sources.

Apart from perceived direction, the perceived distance when using a source with controlled directivity has also been investigated (Laitinen et al. 2015; Wendt et al. 2016). Wendt and colleagues described the relation between perceived distance and beam-pattern control such as beam width (i.e., order) or the angle between a pair of symmetric third-order beams.

Wendt and coworkers (2017) provided a collection of formal listening experiments with the physical IKO's spherical harmonic beamforming in a real room. These experiments test (1) the localization of static beam directions, (2) the localization of time-variant beam steering using different sounds, and (3) whether "sculptural" compositional categories can be distinguished based on spatial impressions, as opposed to impressions based on monophonic playback.

From Wendt et al.'s second experiment, we can present further results demonstrating that beamforming from the physical IKO in a physical room (which is not ideal) is able to influence the impression of distance in the case of time-varying beam-pattern control (for details of the acoustical properties and the exact setup positions of the experiment, cf. Wendt et al. 2017). Fifteen listeners with experience in auditing spatial audio, drawn from IEM's expert listening panel, took part in the experiment, in which each listener gave responses for two listening positions.

The conditions consisted of 5 sec of pink-noise bursts and a 5-sec, time-varying beam-pattern control. Subjects were asked to mark the position of the evoked auditory object in time steps of 0.5 sec, using ten controllable dots on a graphical interface showing the layout of both the room and IKO (cf. Figure 7). Each of the dots could be moved by mouse and flashed at the corresponding moment of playback. Listeners could repeat the playback until they were satisfied with the match of their response and what they perceived.

Binaural renderings using vIKO (Zaunschirm, Frank, and Zotter 2016) are available online: http://phaidra.kug.ac.at/o:37710, http://phaidra.kug.ac.at/o:37712 for the two beam-pattern-control conditions at position 1, and http://phaidra.kug.ac.at/o:37711, http://phaidra.kug.ac.at/o:37711 for position 2.

Figure 7 shows the mean results for each time step for two beam-pattern control conditions: (1) left–right amplitude panning from a beam aiming toward the left to a beam aiming toward the right, (2) distance panning for beam steering to the back wall (i.e., direct sound at listening position 1). The distance panning gradually changes the order of the beam pattern from third order to zeroth order and back again to third order. Distance panning works more clearly at position 1, but also affects the perceived location at position 2. The result for left–right amplitude-panned beam pairs is perceivable from both listening positions and indicates the feasibility of lateral distance control for auditory events created by the IKO.

Part II: Beamformer Theory and Control of the IKO

Based on the technical background provided in Part I, we now look to a deeper understanding of the working principles behind beams formed with compact spherical loudspeaker arrays such as the IKO. It explains the governing acoustic equations in general, and for the IKO as a particular case. A comprehensive approach is outlined for the design

Figure 7. Geometry of the experiment, showing the listening and IKO positions in our lecture room for listening positions 1 (a) and 2 (b). Dark gray circles indicate mean localized positions for distance panning, and light gray squares the positions for left-right panning. Marker size increases with time.

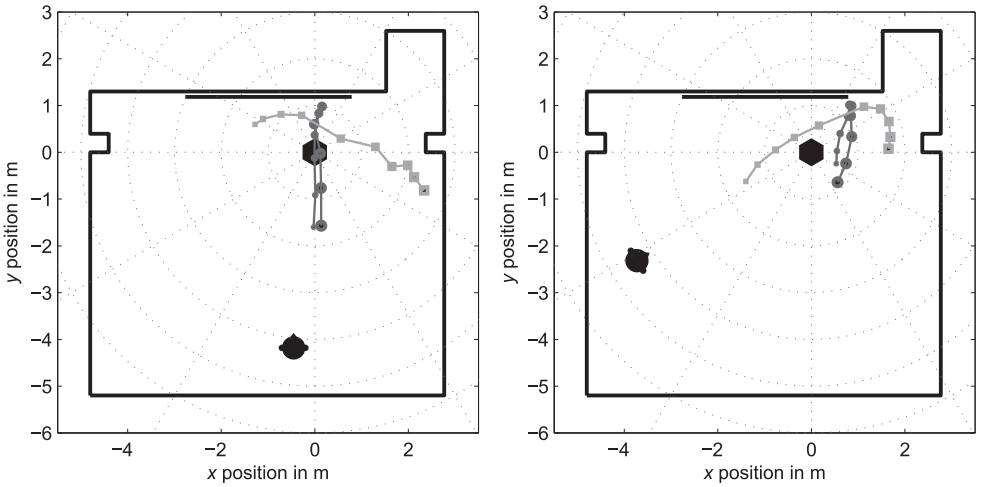


Figure 7

Figure 8. Spherical harmonic patterns up to third order (a). Specific surface vibration patterns (b) can synthesize spherical harmonic beam patterns in the far field (cap model).

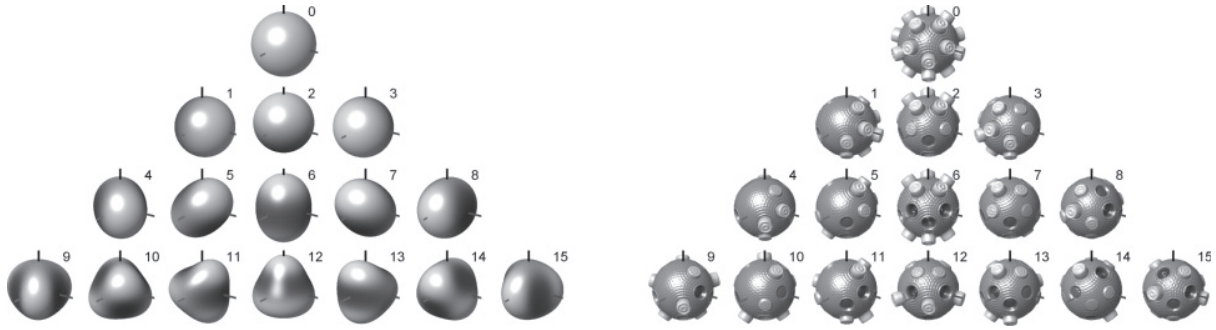


Figure 8

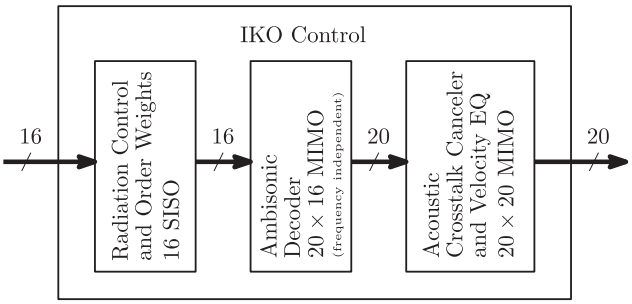
and verification of the filters required to configure the DAW plug-ins. A novel limiting criterion is introduced to safely operate the array by ensuring a limited loudspeaker excursion. The MIMO system design presented here is based on laser Doppler vibrometry measurements. The resulting far-field beam pattern is examined based on microphone-array measurements at a finite distance, and the measured data are extrapolated to the far field to verify the design.

Beamforming with the IKO

The key to controlling focused sound beams with the IKO is the ability to control the sound particle velocity on its surface in the shape of spherical harmonics; see Figure 8a.

The IKO houses 20 passive loudspeakers that are mounted into its rigid faces (shown in Figure 1). Because all loudspeakers of the IKO share one common enclosure volume, the motion of the loudspeaker

Figure 9. MIMO synthesis of spherical harmonic beam patterns.



cones is acoustically coupled. Directivity pattern synthesis requires individual control of the cone motions, however, so that a MIMO crosstalk canceler is needed. The vibrometry-based identification of the canceler is described in the section "Measurement and Control of Loudspeaker Cone Velocity," see the rightmost block of the schema in Figure 9.

According to the equations of sound radiation, any surface velocity vibration pattern in the exact shape of an individual spherical harmonic (cf. Figure 8a) propagates to a sound-pressure pattern of the same shape at any radius. The pattern only undergoes a radius- and frequency-dependent change of magnitude and phase, obeying a well-defined frequency response for each order n of spherical harmonic (Zotter 2009). In the far field, this is

$$b_n(k\mathrm{R}) = \frac{\rho c\, \mathrm{i}^n}{k\, h_n'^{(2)}(k\mathrm{R})}, \qquad (1)$$

where ρ is the density of air (1.2 kg/m^3), c is the speed of sound (343 m/sec), and i is the imaginary unit. The wave number $k = 2\pi f/c$ is defined by the frequency f, and $h_n'^{(2)}(k\mathrm{R})$ is the derivative of the nth-order spherical Hankel function of the second kind that describes radiation for Fourier representations with a positively signed exponent $e^{\mathrm{i}2\pi f t}$ (Zotter 2009). The effective acoustical radius of the IKO is R = 28.5 cm.

Complicated surface vibration patterns are smoothed out as the sound is radiated to the far field. Accordingly, signals decoded to high-order patterns are strongly attenuated, particularly at low frequencies (see Figure 10). The remaining low-order patterns—decoded to the loudspeakers (as seen in Figure 8b) by the signal processing block in the middle of Figure 9—can be equalized by *far-field radiation control* to compensate for the amplitude and phase changes they undergo when radiated (Zotter and Noisternig 2007; Pomberger 2008; Kerscher 2010). This step consists of single-input, single-output filters (SISO) and is accomplished by the leftmost block in Figure 9. This enables one to compose far-field beam patterns out of superimposed low-order spherical harmonics.

Figure 10. Surface vibration patterns on a sphere are radiated to sound pressure in the far field, with frequency responses depending on the spherical harmonic order n. *The diagram shows these responses for a sphere of the radius* R = *28.5 cm. The far-field sound pressure is characterized by the increasingly strong attenuation of components of high orders and low frequencies, yielding (*n + *1)th-order high-pass slopes, cf. Equation 1.*

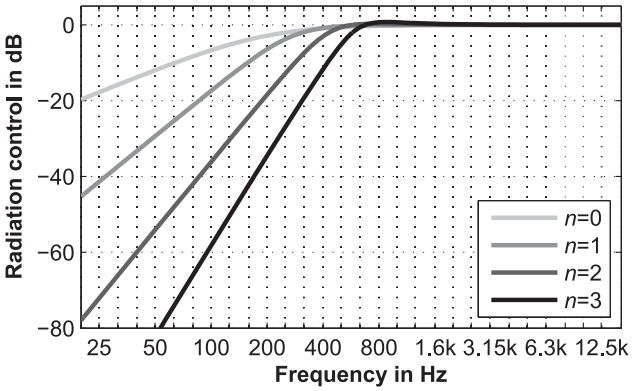


Figure 10

Given a suitable control system achieving the control of the IKO's surface velocity and radiation, the composition of far-field beam patterns in terms of spherical harmonics works using the same tools as for arranging sounds in higher-order Ambisonics, i.e., Ambisonic encoding. Figure 11 demonstrates that, for an aiming of a spherical harmonic beam towards the angle 0° on the horizon, the effort to develop elaborated systems pays off: Narrower beams can be achieved maintaining a more consistent shape over a larger frequency range. Figure 11a shows a beam pattern achieved by plain max-r_E Ambisonic amplitude panning (Zotter, Pomberger, and Schmeder 2008; see also Daniel, Rault, and Polack 1998 for details on max-r_E weighting), and Figure 11b displays a system we designed in 2014 (cf. Lösler and Zotter 2015). To avoid audible distortions due to overload at low frequencies, this design entailed the necessity of a low-frequency amendment by crossing over to an omnidirectional subwoofer mode.

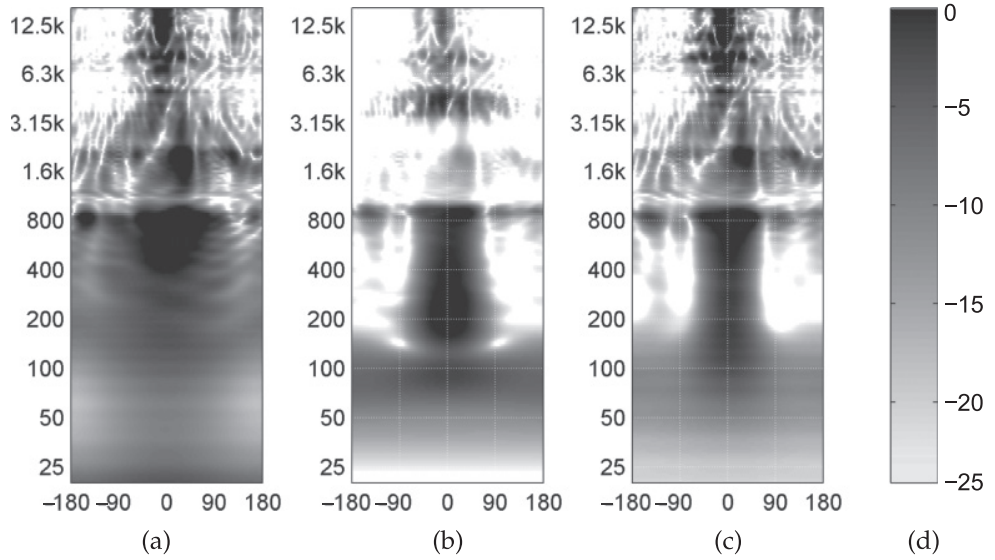


Figure 11. Beam pattern of a horizontal beam, with magnitude in dB (grayscale) over polar angle and frequency for spherical harmonic beamforming with the IKO using systems with plain max-r_E Ambisonic amplitude panning (a); a version from 2014 of radiation control including EQ by ear and MIMO acoustic crosstalk cancellation (b); and the new limited-excursion design without MIMO acoustic crosstalk cancellation but EQ for the active loudspeaker velocity (c). Magnitude levels indicated as levels of gray (d).

Later in this article, the section "Limiting the Loudspeaker Cone Excursion" will present our new design method recognizing excursion as a more reasonable physical limitation than the white-noise gain constraint, which we had previously used and that was adopted from microphone array theory. Even without a MIMO crosstalk canceler for the IKO's loudspeaker cones, this concept achieves beams that are more focused than in our previous design, whose rough equalization by ear obviously led to a lack of energy above 800 Hz (compare the graphs in Figures 11b and 11c).

Desired max-r_E Beam Patterns

Far-field beam-pattern synthesis by the IKO uses the same description as the angular amplitude patterns in higher-order Ambisonics. The max-r_E beam patterns that will be used here turned out to exhibit sufficiently high side-lobe attenuation while maintaining a narrow main lobe (Daniel, Rault, and Polack 1998). On-axis equalized max-r_E beams of the orders $i \leq \mathrm{N}$ are shown in Figure 12 and described in earlier publications (Zotter and Frank 2012; Lösler and Zotter 2015), giving the equation

$$g_i(\boldsymbol{\theta}) = \sum_{n=0}^{\mathrm{N}} \sum_{m=-n}^{n} Y_n^m(\boldsymbol{\theta})\, w_{n,i}\, Y_n^m(\boldsymbol{\theta}_c), \qquad (2)$$

$$w_{n,i} = \begin{cases} \dfrac{P_n\left(\cos\left(\frac{137.9^\circ}{i+1.51}\right)\right)}{\sum_{n=0}^{i}(2n+1)\, P_n\left(\cos\left(\frac{137.9^\circ}{i+1.51}\right)\right)} & \text{for } n \leq i \\ 0 & \text{elsewhere,} \end{cases}$$

where the $Y_n^m(\boldsymbol{\theta})$ are the fully orthonormal spherical harmonics as depicted in Figure 8, and the $P_n(\cdot)$ are nth-order Legendre polynomials (cf. Zotter and Frank 2012). The two direction vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_c$ denote the observed direction of radiation and the adjustable beam direction, respectively. The controllable Ambisonics order is considered to be limited by N, and the weights $w_{n,i}$ are the max-r_E order weights.

The Ambisonic encoding shown in Figure 3 corresponds to the distribution of a single-channel signal to $(\mathrm{N}+1)^2$ channels using the spherical harmonics $Y_n^m(\boldsymbol{\theta}_c)$ evaluated at the beam direction $\boldsymbol{\theta}_c$ as weights, as suggested by the rightmost term in Equation 2. The IKO control system has the task of representing the two leftmost terms, $Y_n^m(\boldsymbol{\theta})$ and

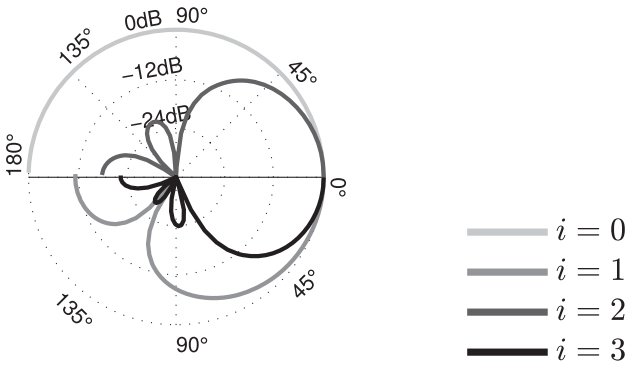


Figure 12. Spherical harmonic max-r_E beam patterns of orders $i = 0, 1, 2, 3$, yielding rotationally symmetrical directivity patterns, which could be drawn as balloon diagrams in three dimensions. Here, the polar diagram shows semicircular generatrix curves in alternation on the interval [0°, ± 180°] to maintain quantitative legibility.

$w_{n,i}$, by achieving the best possible synthesis of all the controllable $(N+1)^2$ max-r_E-weighted spherical harmonics in the far field.

The highest controllable order N depends on the number of loudspeakers L, with $L \geq (N+1)^2$. The highest-order pattern for the $L = 20$ of the IKO is $g_3(\boldsymbol{\theta})$ and its synthesis is difficult to accomplish at low frequencies. Instead, reasonable processing produces a sequence of increasingly focused beam patterns $g_0(\boldsymbol{\theta})$, $g_1(\boldsymbol{\theta})$, $g_2(\boldsymbol{\theta})$, $g_3(\boldsymbol{\theta})$ over frequency.

Cap Model of Surface Vibration

A unity-gain velocity excited by the loudspeaker cone sitting at the direction $\boldsymbol{\theta}_l$ can be modeled as a spherical cap of the aperture angle α. For a variable direction of observation, this is expressed as a unit step function $u(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}_l - \cos(\alpha/2))$. Its contribution to each spherical harmonic is defined by the transform

$$\nu_{nm}^{(l)}|_{\mathrm{R}} = \int u\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}_l - \cos\frac{\alpha}{2}\right) \mathrm{Y}_n^m(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta},$$

(cf. Zotter, Sontacchi, and Höldrich 2007), yielding

$$\nu_{nm}^{(l)}|_{\mathrm{R}} = a_n\, \mathrm{Y}_n^m(\boldsymbol{\theta}_l),$$

$$a_n = \begin{cases} P_{n-1}(\cos\frac{\alpha}{2}) - \cos\frac{\alpha}{2}\, P_n(\cos\frac{\alpha}{2}) & n > 0 \\ 1 - \cos(\frac{\alpha}{2}) & n = 0. \end{cases}$$

A weighted superposition of all the IKO's loudspeaker cones, assuming their velocities are given by the weights v_l, is written as

$$\nu_{nm}|_{\mathrm{R}} = a_n \sum_{l=1}^{\mathrm{L}} \mathrm{Y}_n^m(\boldsymbol{\theta}_l)\, v_l.$$

We can stack the $(N+1)^2$ coefficients of spherical harmonics $\nu_{nm}|_{\mathrm{R}}$ into a vector $\boldsymbol{\nu}_{\mathrm{N}} = [\nu_{nm}]$ and the L loudspeaker velocities of the IKO into another vector $\boldsymbol{v} = [v_l]$. The a_n weights, written as vector $\boldsymbol{a}_{\mathrm{N}} = [a_n]$, and the spherical harmonics up to the order N sampled at the 20 loudspeakers, written as matrix $\boldsymbol{Y}_{\mathrm{N}} = [\mathrm{Y}_n^m(\boldsymbol{\theta}_l)]_{nm}^{l}$, permit us to express the matrix equation as $\boldsymbol{\nu}_{\mathrm{N}}|_{\mathrm{R}} = \mathrm{diag}\{\boldsymbol{a}_{\mathrm{N}}\}\boldsymbol{Y}_{\mathrm{N}}\,\boldsymbol{v}$ whose least-squares inverse

$$\boldsymbol{v} = \underbrace{\boldsymbol{Y}_{\mathrm{N}}^{\mathrm{T}}\left(\boldsymbol{Y}_{\mathrm{N}}\boldsymbol{Y}_{\mathrm{N}}^{\mathrm{T}}\right)^{-1}}_{:=\boldsymbol{D}_{\mathrm{N}}} \mathrm{diag}\{\boldsymbol{a}_{\mathrm{N}}\}^{-1}\ \boldsymbol{\nu}_{\mathrm{N}}|_{\mathrm{R}}$$

yields suitable loudspeaker velocities. Expressed in its scalar form, with the decoder coefficients $\boldsymbol{D}_{\mathrm{N}} = [d_{nm}^{(l)}]$, the loudspeaker velocities v_l producing the coefficients $\nu_{nm}|_{\mathrm{R}}$ are

$$v_l = \sum_{n=0}^{\mathrm{N}} \sum_{m=-n}^{n} \frac{d_{nm}^{(l)}}{a_n}\, \nu_{nm}|_{\mathrm{R}}. \qquad (3)$$

Radiation Control

As specified by $b_n(k\mathrm{R})$ in the frequency domain (cf. Equation 1 and Figure 10), the spherical harmonic coefficient of the surface velocity $\nu_{nm}|_{\mathrm{R}}$ radiates into the far field, yielding the sound-pressure coefficient

$$\psi_{nm} = b_n(k\mathrm{R})\, \nu_{nm}|_{\mathrm{R}}.$$

The aim is to control this coefficient to obtain a far-field beam pattern as in Equation 2

$$\psi_{nm} = \mathrm{Y}_n^m(\boldsymbol{\theta}_c)\, w_{n,i},$$

so we invert the equation to obtain

$$\nu_{nm}|_{\mathrm{R}}(f, \boldsymbol{\theta}_c) = \frac{w_{n,i}}{b_n(k\mathrm{R})}\, \mathrm{Y}_n^m(\boldsymbol{\theta}_c).$$

Insertion into Equation 3 yields the loudspeaker velocities required to produce the desired max-r_E

beam pattern $g_i(\boldsymbol{\theta})$, here with $i = 0, 1, 2, 3$,

$$v_l^{(i)}(f, \boldsymbol{\theta}_c) = \sum_{n=0}^{\mathrm{N}} \sum_{m=-n}^{n} \frac{d_{nm}^{(l)}}{a_n} \frac{w_{n,i}}{b_n(k\mathrm{R})} Y_n^m(\boldsymbol{\theta}_c).$$

Figure 10 shows that the inverse of $b_n(k\mathrm{R})$ can require unrealistic bass boosts to compensate for attenuation of higher orders. For spherical microphone arrays, realistic implementations consider filtering into successive frequency bands $H_i(f)$, in which only an increasingly focused beam pattern $g_i(\boldsymbol{\theta})$ is synthesized with $i = 0, 1, 2, 3$ (Lösler and Zotter 2015). Summed over these bands, the required loudspeaker velocities become

$$v_l(f, \boldsymbol{\theta}_c) = \sum_{n=0}^{\mathrm{N}} \sum_{m=-n}^{n} d_{nm}^{(l)} \sum_{i=0}^{\mathrm{N}} \underbrace{\frac{H_i(f)}{a_n} \frac{w_{n,i}}{b_n(k\mathrm{R})}}_{:=H_{i,n}(f)} Y_n^m(\boldsymbol{\theta}_c). \quad (4)$$

The radiation control filters obtained in this way, $H_{i,n}(f)$, depend on the spherical harmonic order n and the synthesized beam order i. The question is how to design the filters $H_i(f)$ they contain to comply with physical limitations.

Limiting the Loudspeaker Cone Excursion

In microphone arrays, white-noise gain limitation prevents self-noise amplification (Lösler and Zotter 2015). For loudspeaker arrays, the limiting relates instead to a maximum linear transducer excursion $|x_l| \leq \mathrm{x}_{\max}$. Excursion is defined by integrating velocity over time. Accordingly, we formulate the constraint in the frequency domain

$$\max_{l,\boldsymbol{\theta}_c} |x_l(f, \boldsymbol{\theta}_c)| \leq \mathrm{x}_{\max}$$

using $x_l(f, \boldsymbol{\theta}_c) = v_l(f, \boldsymbol{\theta}_c)/\mathrm{i}2\pi f$.

The original radiation control filters $1/b_n(k\mathrm{R})$ exhibit slopes proportional to $1/f^{n+1}$. By the additional factor $1/\mathrm{i}2\pi f$, slopes for unlimited excursion are proportional to $1/f^{n+2}$. Hence, high-pass filters stabilizing the individual radiation control filters by enforcing a constant excursion limit must at least be proportional to f^{n+2}. What is more, whenever such an excursion limit takes effect, the magnitude of the corresponding radiation pattern will vanish in the far field. To exclusively drive excursions producing audible sounds, economic use of excursion requires limitation filters of slopes proportional to f^{n+3}, at least. We define the following filter bank using zero-phase high- and low-pass filters:

$$\hat{H}_0(f) = \frac{(f/f_0)^3}{1+(f/f_0)^3} \frac{1}{1+(f/f_1)^4},$$

$$\hat{H}_1(f) = \frac{(f/f_1)^4}{1+(f/f_1)^4} \frac{1}{1+(f/f_2)^5},$$

$$\hat{H}_2(f) = \frac{(f/f_2)^5}{1+(f/f_2)^5} \frac{1}{1+(f/f_3)^6},$$

$$\hat{H}_3(f) = \frac{(f/f_3)^6}{1+(f/f_3)^6}.$$

To make these filters complementary in amplitude to an overall high-pass filter

$$H_{\mathrm{sum}}(f) = \frac{(f/f_0)^3}{1+(f/f_0)^3},$$

they are normalized using

$$H_i(f) = \frac{(f/f_0)^3}{1+(f/f_0)^3} \frac{\hat{H}_i(f)}{\sum_{i=0}^{\mathrm{N}} \hat{H}_i(f)}.$$

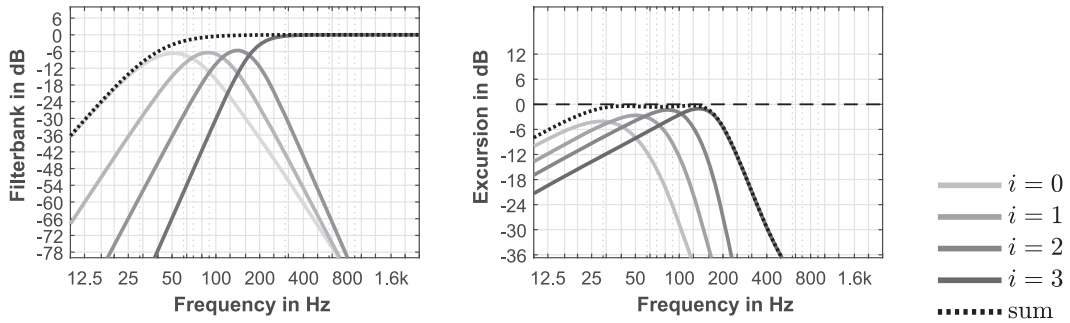
Inserted into Equation 4, with suitable cut-on frequencies f_i, the filters yield limited excursion curves as show in Figure 13b, where the excursion was normalized by the excursion reached at 40 Hz in omnidirectional radiation mode (dashed line).

Measurement and Control of Loudspeaker Cone Velocity

This section deals with measuring the voltage-to-velocity transfer functions of the IKO's loudspeakers, including acoustic coupling (crosstalk) between active and passive movements of the loudspeaker cones. The measured matrix is used to design an equalized and crosstalk-canceled control system for loudspeaker velocities. The voltage-to-velocity transfer functions of the 20 loudspeakers were

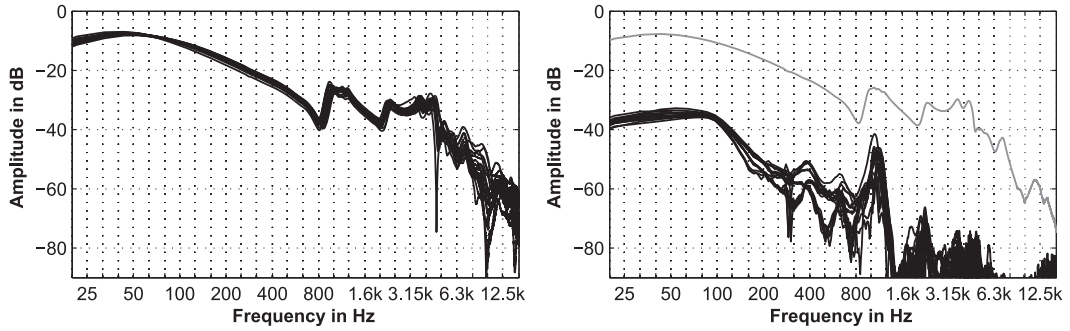
Figure 13. Filter bank $\mathrm{H_i(f)}$ *and overall response using suitably chosen cut-on frequencies* $[\mathrm{f_i}]^\mathrm{T} = [40, 70, 113, 173]$ *Hz (a), and the resulting limited excursion normalized at 40 Hz (b).*

Figure 14. Active voltage-to-velocity responses of all 20 loudspeakers (a), and 19 passive responses to excitation voltages at loudspeaker 1 out of $\mathbf{T(f)} = [\mathrm{t_{ij}(f)}]$ *(b).*



(a) (b)

Figure 13



(a) (b)

Figure 14

measured using the exponential sine-sweep method (Farina 2000) and a laser Doppler vibrometer along the cone axis, approximately 24 cm away from each loudspeaker (as shown in Figure 5a). All measured impulse responses were cropped to 4,096 samples at a 44.1-kHz sample rate. Figure 14 shows some frequency responses.

With the transfer-function matrix in the frequency domain $\boldsymbol{T}$, the output velocities $\boldsymbol{v} = [v_1, \ldots, v_\mathrm{L}]^T$ caused by the input voltages $\boldsymbol{u} = [u_1, \ldots, u_\mathrm{L}]^T$ are calculated as

$$\boldsymbol{v} = \boldsymbol{T}\,\boldsymbol{u}.$$

The frequency dependency is omitted from this notation to maintain simplicity. Given that $\boldsymbol{T}$ is invertible, decoupled cone velocities $\boldsymbol{v}$ are controlled by the voltages

$$\boldsymbol{u} = \boldsymbol{T}^{-1}\,\boldsymbol{v},$$

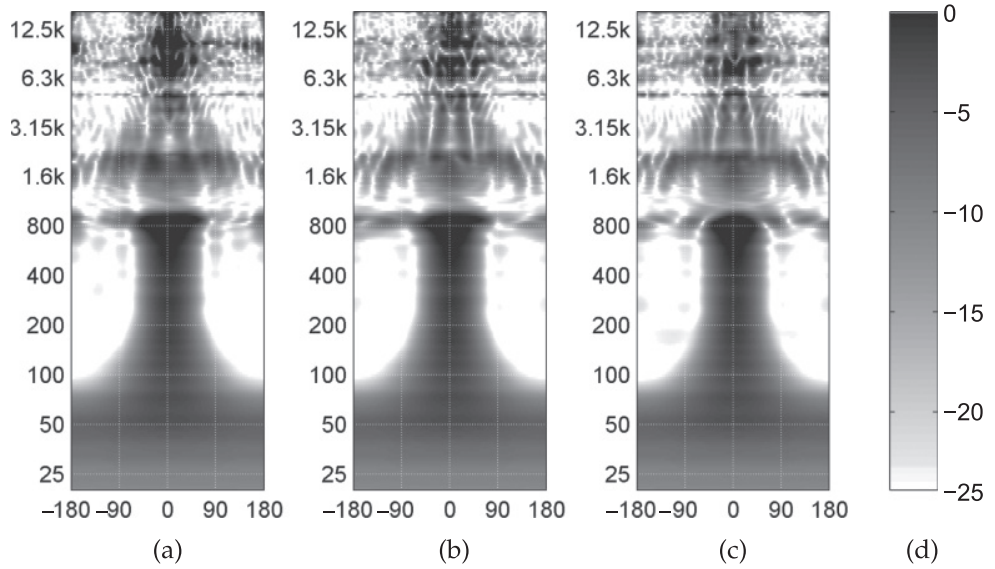
and one can insert Equation 4 for beamforming. To keep the corresponding impulse responses short and easy to window in the time domain, a regularized inverse $\boldsymbol{T}^\mathrm{H}(\boldsymbol{T}\boldsymbol{T}^\mathrm{H} + a\,\frac{\mathrm{tr}\{\boldsymbol{T}\boldsymbol{T}^\mathrm{H}\}}{\mathrm{L}}\boldsymbol{I})^{-1}$ was used, with $\mathrm{L} = 20$. The regularization was set to $a = 0.1$.

The 20×20 voltage-to-velocity impulse responses are available at http://phaidra.kug.ac.at/o:37716.

Verification by Sound Pressure Measurements

Verification of whether the far-field beam pattern complies with the desired max-r_E beam pattern

Figure 15. Horizontal cuts through the on-axis direction of far-field beam patterns of IKO, using a MIMO crosstalk canceler and limited-excursion radiation control. Using polar angle and frequency as axes, the diagrams show dB values for three beam directions: on-axis direction of loudspeaker 6 (a); direction between loudspeakers 6, 11, and 15 (b); and directions between loudspeakers 6 and 11 (c). Magnitude levels are indicated as levels of gray (d).



has been achieved by microphone measurements surrounding the IKO, as shown in Figure 5b (cf. also Zotter and Bank 2012). The impulse responses of the loudspeaker-to-microphone paths were also measured using the exponential sine-sweep technique and windowing to 320 samples. A sine-square fade-in of 20 samples was used before the first impulse and an 80-sample fade-out at the very end.

The 648 × 20 voltage-to-sound-pressure impulse responses are available at http://phaidra.kug.ac.at/o:37715.

In the frequency domain, the sound-pressure sample $p_j|_{75\text{ cm}}$ received at the jth microphone due to the driving voltages u_l from each of the loudspeakers is described by the transfer paths $g_{lj}(f)$ of a MIMO system,

$$p_j|_{75\text{ cm}}(f) = \sum_{l=1}^{\mathrm{L}} g_{lj}(f)\, u_l. \tag{5}$$

The least-square-error inverse

$$\boldsymbol{C} = (\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}})^{-1}\,\boldsymbol{Y}^{\mathrm{T}} = [c_{nm}^{(j)}]$$

of $\boldsymbol{Y} = [Y_{nm}(\boldsymbol{\theta}_j)]$ (the spherical harmonics sampled at the microphone positions $\boldsymbol{\theta}_j$) permits decomposition into coefficients of spherical harmonics $\psi_{nm}|_{75\text{ cm}}$ up to $n \le 17$ with the given measurement setup.

The far-field sound pressure is calculated from this decomposition at a desired cross section

$$p_{\mathrm{ff}}(f,\boldsymbol{\theta}) = \sum_{n=0}^{17}\sum_{m=-n}^{n} \frac{\mathrm{i}^{n+1}\, Y_n^m(\boldsymbol{\theta})}{k\, h_n(k75\text{ cm})} \sum_{j=1}^{648} c_{nm}^{(j)}\, p_j|_{75\text{ cm}}(f).$$

Figure 15 shows a cross section centered on beams with different orientations after inserting the IKO control into Equation 5. The new beam patterns are more constant and narrow in the range from 150 Hz to 800 Hz than they were with earlier approaches.

At 200 Hz, the previous approaches in Figure 11 yield beams reaching an attenuation of approximately −6 dB from their maximum at angles between ±135° (Figure 11a), ±60° (Figure 11b), ±50° (Figure 11c), and the proposed filter design in Figure 15 reaches this value at ±45°. At 100 Hz, the proposed design achieves a ±70-degree width for a −6 dB attenuation from its maximum. With the previous designs, only the design in Figure 11c achieved focus at all to ±80°, but it was not able to maintain the amplitude below 100 Hz.

The new beam patterns in Figure 15 become roughly omnidirectional below 100 Hz, and above 800 Hz the inherent spatial aliasing counteracts a smooth beam pattern. A notch around 1 kHz appears, most probably because of a mismatch of

the loudspeaker cone vibration from an ideally rigid shape. Around the frequencies 1.6 kHz, 5 kHz, and 8 kHz, the IKO seems to lose its directivity. This is probably caused by modal breakup or interior modes of the IKO.

Conclusion

We presented the IKO, a new computer music instrument utilizing superdirective spherical harmonic beamforming to orchestrate the wall reflections in a room. We could outline its use by free, ready-to-use DAW plug-ins enabling its spherical harmonic beamforming in real time, and its use as a virtualized instrument (vIKO) that is freely available together with illustrative binaural renderings. We gave a precise description of our IKO prototype and basic concert setups that were used in the past.

To describe perceptual aspects of the spherical harmonic beamforming with the IKO, we reviewed previous experiments and showed new results indicating that the IKO allows one to control the direction as well as the distance impression of the synthesized sound objects.

We summarized the spherical harmonic beamforming theory of compact spherical arrays and presented a simple way of defining a bank of linear-phase limitation filters that suppress side lobes in each of its frequency bands. We were able to outline constraints that are relevant for compact spherical loudspeaker arrays, since the more common white-noise-gain limitations, as applicable to spherical microphone arrays, become meaningless in this context.

Finally, we presented a practical study to measure responses to design an entire multiple-input, multiple-output control filter set. It is based on laser Doppler vibrometry measurements for a clean control of the IKO's loudspeaker cone velocities, with crosstalk cancelled; excursion-limited analytic filter design, suppressing side lobes, for radiation control; and measurements verifying the synthesized radiation patterns by using far-field extrapolated measurements with a spherical microphone array. All measurement data are made available to support reproducible research.

Acknowledgments

This work was funded by the Austrian Science Fund (FWF), project no. AR 328-G21, "Orchestrating Space by Icosahedral Loudspeaker." We would like to thank *Computer Music Journal*'s anonymous reviewers and Editor, and Frank Schultz of sonible, for their valuable comments on our manuscript.

References

Avizienis, R., et al. 2006. "A Compact 120 Independent Element Spherical Loudspeaker Array with Programmable Radiation Patterns." In *Proceedings of the 120th Audio Engineering Society Convention*. Available online at www.aes.org/e-lib/browse.cfm?elib=13587 (subscription required). Accessed April 2017.

Bennett, M. B., and D. T. Blackstock. 1975. "Parametric Array in Air." *Journal of the Acoustical Society of America* 57(3):562–568.

Bloch, A., R. Medhurst, and S. Pool. 1953. "A New Approach to the Design of Super-Directive Aerial Arrays." *Proceedings of the IEE-Part III: Radio and Communication Engineering* 100(67):303–314.

Boulez, P. 1983. "L'Acoustique et la musique contemporaine: Introduction." In *Revue d'acoustique: Congrès international d'acoustique*, pp. 213–216.

Brandstein, M., and D. Ward. 2001. *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin: Springer.

Butler, J. L., and S. L. Ehrlich. 1977. "Superdirective Spherical Radiator." *Journal of the Acoustical Society of America* 61(6):1427–1431.

Caussé, R., J. Bresciani, and O. Warusfel. 1992. "Radiation of Musical Instruments and Control of Reproduction with Loudspeakers." In *Proceedings of the International Symposium on Musical Acoustics*, pp. 67–70.

Cook, P., et al. 1998. "N>>2: Multi-Speaker Display Systems for Virtual Reality and Spatial Audio Projection." In *Proceedings of the International Conference on Auditory Display*. Available online at www.icad.org/Proceedings. Accessed 21 March 2017.

Croft, J. J., and J. O. Norris. 2003. "Theory, History, and the Advancement of Parametric Loudspeakers: A Technology Overview." Technical Report 98-10006-1100 Rev. E. San Diego, California: American Technology Corporation.

Daniel, J., J.-B. Rault, and J.-D. Polack. 1998. "Ambisonics Encoding of Other Audio Formats for Multiple Listening

Conditions." In *Proceedings of the 105th Audio Engineering Society Convention*. Available online at www.aes.org/e-lib/browse.cfm?elib=8385 (subscription required). Accessed April 2017.

Elko, G. W. 2000. "Superdirectional Microphone Arrays." In J. Benesty and S. L. Gay, eds. *Acoustic Signal Processing for Telecommunication*. Berlin: Kluwer.

Farina, A. 2000. "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique." In *Proceedings of the 108th Audio Engineering Society Convention*. Available online at www.aes.org/e-lib/browse.cfm?elib=10211 (subscription required). Accessed April 2017.

Frank, M., G. K. Sharma, and F. Zotter. 2015. "What We Already Know about Spatialization with Compact Spherical Arrays as Variable-Directivity Loudspeakers." Paper presented at the inSonic Conference, 26–28 November, Karlsruhe, Germany. Available online at iem.kug.ac.at/fileadmin/media/osil/2015_FrankEtAl_inSonic_WhatWeAlreadyKnowAboutSpatialization WithCompactSphericalArraysAsVariabledirectivity Loudspeakers.pdf. Accessed March 2017.

Kassakian, P. 2005. "Magnitude Least-Squares Fitting via Semidefinite Programming with Applications to Beamforming and Multidimensional Filter Design." In *Proceedings of the IEEE International Conference onAcoustics, Speech, and Signal Processing*. Available online at ieeexplore.ieee.org/document/1415644 (subscription required). Accessed May 2017.

Kassakian, P. 2006. "Convex Approximation with Applications in Magnitude Filter Design and Beamforming." PhD dissertation, Electrical Engineering and Computer Science, University of California, Berkeley. Available online at www2.eecs.berkeley.edu /Pubs/TechRpts/2006/EECS-2006-64.pdf. Accessed March 2017.

Kassakian, P., and D. Wessel. 2003. "Design of Low-Order Filters for Radiation Synthesis." In *Proceedings of the 115th Audio Engineering Society Convention*. Available online at www.aes.org/e-lib/browse.cfm?elib=12448 (subscription required). Accessed April 2017.

Kassakian, P., and D. Wessel. 2004. "Characterization of Spherical Loudspeaker Arrays." In *Proceedings of the 117th Audio Engineering Society Convention*. Available online at www.aes.org/e-lib/browse.cfm?elib=12940 (subscription required). Accessed April 2017.

Kerscher, M. 2010. *Compact Spherical Loudspeaker Array, Implementation of a System for Variable Sound Radiation*. Saarbrücken: VDM.

Kronlachner, M. 2014. "Plug-in Suite for Mastering the Production and Playback in Surround Sound and Ambisonics." Paper presented at the 136th Audio Engineering Society Convention Student Design Competition, 28 April 2014, Berlin. Available online at www.matthiaskronlachner.com/wp-content/uploads /2013/01/kronlachner_aes_studentdesigncompetition _2014.pdf. Accessed March 2017.

Laitinen, M.-V., et al. 2015. "Controlling the Perceived Distance of an Auditory Object by Manipulation of Loudspeaker Directivity." *Journal of the Acoustical Society of America* 137(6):462–468.

Lindemann, W. 1986. "Extension of a Binaural Cross-Correlation Model by Contralateral Inhibition: I. Simulation of Lateralization for Stationary Signals." *Journal of the Acoustical Society of America* 80(6):1608–1622.

Lösler, S., and F. Zotter. 2015. "Comprehensive Radial Filter Design for Practical Higher-Order Ambisonic Recording." In *Fortschritte der Akustik: Tagungsband der deutschen Arbeitsgemeinschaft für Akustik*, pp. 452–455.

Miranda, L., D. Cabrera, and K. Stewart. 2013. "A Concentric Compact Spherical Microphone and Loudspeaker Array for Acoustical Measurements." In *Proceedings of the 135th Audio Engineering Society Convention*. Available online at www.aes.org /e-lib/browse.cfm?elib=16986 (subscription required). Accessed April 2017.

Misdariis, N., et al. 2001. "Radiation Control on Multi-Loudspeaker Device: La Timée." In *Proceedings of the International Computer Music Conference*, pp. 306–309.

Morgenstern, H., B. Rafaely, and F. Zotter. 2015. "Theory and Investigation of Acoustic Multiple-Input Multiple-Output Systems Based on Spherical Arrays in a Room." *Journal of the Acoustical Society of America* 138(5):2998–3009.

Morgenstern, H., F. Zotter, and B. Rafaely. 2012. "Joint Spherical Beam Forming for Directional Analysis of Reflections in Rooms." *Journal of the Acoustical Society of America* 131(4):3207–3207.

Pasqual, A. M. 2010. "Sound Directivity Control in a 3-D Space by a Compact Spherical Loudspeaker Array." PhD dissertation, University of Campinas, Faculty of Mechanical Engineering, Campinas, Brazil.

Pasqual, A. M., J. R. Arruda, and P. Herzog. 2010. "Application of Acoustic Radiation Modes in the Directivity Control of a Spherical Loudspeaker Array." *Acta Acustica United with Acustica* 96(1):32–42.

Pasqual, A. M., P. Herzog, and J. R. Arruda. 2010. "Theoretical and Experimental Analysis of the Behavior of a Compact Spherical Loudspeaker Array for Directivity

Control." *Journal of the Acoustical Society of America* 128(6):3478–3488.

Poletti, M. A., T. Betlehem, and T. D. Abhayapala. 2015. "Higher-Order Loudspeakers and Active Compensation for Improved 2D Sound Field Reproduction in Rooms." *Journal of the Audio Engineering Society* 63(1–2):31–45.

Pollow, M. 2014. "Directivity Patterns for Room Acoustical Measurements and Simulations." PhD dissertation, RWTH-Aachen, Institute of Technical Acoustics.

Pollow, M., and G. K. Behler. 2009. "Variable Directivity for Platonic Sound Sources Based on Spherical Harmonics Optimization." *Acta Acustica United with Acustica* 95(6):1082–1092.

Pomberger, H. 2008. "Angular and Radial Directivity Control for Spherical Loudspeaker Arrays." Master's thesis, University of Music and Performing Arts, Institute of Electronic Music and Acoustics, Graz, Austria.

Rafaely, B., and D. Kaykin. 2011. "Optimal Model-Based Beamforming and Independent Steering for Spherical Loudspeaker Arrays." *IEEE Transactions on Audio Speech and Language Processing* 19(7):2234–2238.

Rakerd, B., W. M. Hartmann, and J. Hsu. 2000. "Echo Suppression in the Horizontal and Median Sagittal Planes." *Journal of the Acoustical Society of America* 107(2):1061–1064.

Schelkunoff, S. A. 1943. "A Mathematical Theory of Linear Arrays." *The Bell System Technical Journal* 22(1):80–107.

Sharma, G. K. 2016. "Composing with Sculptural Sound Phenomena in Computer Music." PhD dissertation, University of Music and Performing Arts, Institute of Electronic Music and Acoustics, Graz, Austria.

Sharma, G. K., F. Zotter, and M. Frank. 2014. "Orchestrating Wall Reflections in Space by Icosahedral Loudspeaker: Findings from First Artistic Research Exploration." In *Proceedings of the Joint International Computer Music Conference and the Sound and Music Computing Conference*, pp. 830–835.

Sugibayashi, Y., et al. 2012. "Three-Dimensional Acoustic Sound Field Reproduction Based on Hybrid Combination of Multiple Parametric Loudspeakers and Electrodynamic Subwoofer." *Applied Acoustics* 73(12):1282–1288.

Takumai, S. 2006. "Loudspeaker Array Device and Method for Setting Sound Beam of Loudspeaker Array Device." WIPO patent WO 2,006,001,272 A1, filed 21 June 2005, and granted 5 January 2006.

Trueman, D., et al. 2006. "PLOrk: The Princeton Laptop Orchestra, Year 1." In *Proceedings of the International Computer Music Conference*, pp. 443–450.

Warusfel, O., P. Derogis, and R. Caussé. 1997. "Radiation Synthesis with Digitally Controlled Loudspeakers." In *Proceedings of the 103rd Audio Engineering Society Convention*. Available online at www.aes.org/e-lib/browse.cfm?elib=7202 (subscription required). Accessed April 2017.

Wendt, F., et al. 2016. "Directivity Patterns Controlling the Auditory Distance." In *Proceedings of the International Conference on Digital Audio Effects*, pp. 295–300.

Wendt, F., et al. 2017. "Perception of Spatial Sound Phenomena Created by the Icosahedral Loudspeaker." *Computer Music Journal* 41(1):76–88.

Zaunschirm, M., M. Frank, and F. Zotter. 2016. "An Interactive Virtual Icosahedral Loudspeaker Array." In *Fortschritte der Akustik: Tagungs-CD der deutschen Arbeitsgemeinschaft für Akustik*, pp. 1331–1334.

Zotter, F. 2009. "Analysis and Synthesis of Sound-Radiation with Spherical Arrays." PhD dissertation, University of Music and Performing Arts, Institute of Electronic Music and Acoustics, Graz, Austria.

Zotter, F., and B. Bank. 2012. "Geometric Error Estimation and Compensation in Compact Spherical Loudspeaker Array Calibration." In *Proceedings of the IEEE International Instrumentation and Measurement Technology Conference*, pp. 2710–2715.

Zotter, F., and M. Frank. 2012. "All-Round Ambisonic Panning and Decoding." *Journal of the Audio Engineering Society* 60(10):807–820.

Zotter, F., and M. Frank. 2015. "Investigation of Auditory Objects Caused by Directional Sound Sources in Rooms." *Acta Physica Polonica A* 128(1-A):5–10.

Zotter, F., and R. Höldrich. 2007. "Modeling Radiation Synthesis with Spherical Loudspeaker Arrays." In *Proceedings of the International Congress on Acoustics*. Available online at iem.kug.ac.at/fileadmin/media/iem/altdaten/projekte/publications/paper/modeling_radiation/modeling.pdf. Accessed 21 March 2017.

Zotter, F., and M. Noisternig. 2007. "Near- and Far-Field Beamforming Using Spherical Loudspeaker Arrays." In *Congress of the Alps Adria Acoustics Association*. Available online at iem.kug.ac.at/fileadmin/media/iem/altdaten/projekte/publications/paper/near/near.pdf. Accessed 21 March 2017.

Zotter, F., H. Pomberger, and A. Schmeder. 2008. "Efficient Directivity Pattern Control for Spherical Loudspeaker Arrays." In *Proceedings of the Joint Meeting of the Acoustical Society of America and the European Acoustics Association*. Available online at

iem.kug.ac.at/fileadmin/media/iem/altdaten/projekte/publications/paper/efficient/efficient.pdf. Accessed March 2017.

Zotter, F., and A. Sontacchi. 2007. "Icosahedral Loudspeaker Array." Technical Report IEM Report 39/07. University of Music and Performing Arts, Institute of Electronic Music and Acoustics.

Zotter, F., A. Sontacchi, and R. Höldrich. 2007. "Modeling a Spherical Loudspeaker System as Multipole Source." In *Fortschritte der Akustik: Tagungsband der deutschen Arbeitsgemeinschaft für Akustik*, pp. 221–222.

Zotter, F., et al. 2014. "Preliminary Study on the Perception of Orientation-Changing Directional Sound Sources in Rooms." In *Proceedings of the Forum Acusticum*. Paper presented at the Forum Acusticum, 7–12 September, Krakow, Poland. Available online at ambisonics.iem.at/Members/zotter/2014_zotter_OrientationDirectionalSource.pdf. Accessed March 2017.

3.2 Directivity and Electro-Acoustic Measurements of the IKO

This work was published as:

F. Schultz, **M. Zaunschirm**, and F. Zotter. (2018). Directivity and electro-acoustic measurements of the IKO. *Audio Engineering Society Convention e-Brief 144*, Milano.

Most of the manuscript was written by the first author with periodic contributions from me, as the second author, and the third author. I contributed significantly to the measurements and the design of the IKO's directivity filters.

Audio Engineering Society

Convention e-Brief 444

Presented at the 144th Convention
2018 May 23 – 26, Milan, Italy

Directivity and Electro-Acoustic Measurements of the IKO

Frank Schultz, Markus Zaunschirm, and Franz Zotter

Institute of Electronic Music and Acoustics (IEM), University of Music and Performing Arts Graz (KUG), Austria

Correspondence should be addressed to Frank Schultz (frank.schultz@kug.ac.at)

ABSTRACT

The icosahedral loudspeaker (IKO) as a compact spherical array is capable of 3rd order Ambisonics (TOA) beamforming, and it is used as musical and technical instrument. To develop and verify beamforming with its 20 loudspeakers flush-mounted into the faces of the regular icosahedron, electro-acoustic properties must be measured. We offer a collection of measurement data of IEM's IKO1, IKO2 and IKO3 along with analysis tools to inspect these properties. Multiple-input-multiple-output (MIMO) data comprises: (i) laser vibrometry measurements of the 20x20 transfer functions from driving voltages to loudspeaker velocities, (ii) 20x16 finite impulse responses (FIR) of the TOA decoding filters, and (iii) 648x20 directional impulse responses from driving voltages to radiated sound pressure. With the open data sets, open source code, and resulting directivity patterns, we intend to support reproducible research about beamforming with spherical loudspeaker arrays.

1 Introduction

Compact spherical loudspeaker arrays are capable of grating-lobe free beamforming into all directions, with rotation-invariant beam patterns over several octaves. They are used as technical and musical instruments, such as for (room) acoustic measurement, home entertainment, media installations and computer music. The latter introduced a convenient approach for composing sonic sculptures, also called plastic sound objects [1] by IKO's sound beams exciting room reflections [2].

The regular icosahedral cabinet with 20 faces and 20 individually controlled 6-inch full-band drivers is a good compromise for numerical robust TOA based beamforming utilizing 16 spherical eigenmodes. It provides sufficient low-frequency performance and a reasonably high spatial aliasing frequency, thus exciting the whole audio bandwidth with proper beam control between approximately 100 Hz and 1 kHz [2].

A first prototype IKO1 was built at IEM in 2006, using an edge length 0.345 m and initially 6.5", later 6" drivers, cf. IKO history[1]. To promote technical and artistic research and based on the idea of improved mobility and compactness, the prototypes IKO2 and IKO3—manufactured by the Graz based company sonible utilizing 6" drivers—were acquired in 2016 and 2018 within the scope of the dedicated *IKO by IEM and sonible* cooperation[2]. Compared to IKO1, the newer prototypes IKO2 and IKO3 exhibit more powerful and technically improved transducers along with a smaller cabinet size. IKO3 is slightly larger than IKO2 as a consequence of an improved manufacturing process (edge lengths: 0.288 m for IKO2 vs. 0.294 m for IKO3). IKO1 was utilized in the studies [1, 2] and IKO2 was deployed for [3, 4].

[1] https://iem.kug.ac.at/projects/osil/about-the-iko.html

[2] https://iko.sonible.com

2 Open Source Data and Software

To research, develop, improve, and validate beamforming, electro-acoustic properties were measured for IKO1,2,3. These measurements are continuously updated on our open access digital repositories located at Phaidra[3]. The present engineering brief accompanies the effort to collect directivity dedicated measurements and derived data within a consistent framework and to provide analyzing tools. For ongoing research at IEM, cf. [5, 6], the dedicated open data webpage [4] contains documentation, Phaidra links of the data and analyzing software. For each IKO the SOFA[5] formated data

- 20x20 transfer impulse responses from driving voltages to loudspeaker velocities using laser vibrometry
- 20x16 FIR TOA decoding filters derived from the velocity measurements
- 648x20 (IKO1,2) and 540x20 (IKO3) directional impulse responses from loudspeaker driving voltage to calibrated microphone receiving voltages

are available at the Phaidra link[6]. For analyzing IKO beamforming the open source tool *balloon_holo* is provided, cf. Fig. 1. The software is capable of loading the SOFA data and of interactively inspecting balloon, polar and surface directivity plots. Furthermore, the TOA decoding FIR filters and configuration files are provided for DAW support with the *mcfx_convolver* plugin[7] and the *ambix_encoder_o3* plugin[8].

2.1 Velocity Measurements

The impulse response of each loudspeaker cone's center was measured with a Polytec PDV-100 laser vibrometer. All loudspeakers act on a common volume, which couples their vibrations. Thus, a matrix of 20x20 crosstalk impulse responses, whose diagonal entries contain the active paths, was acquired to design TOA decoding filters including velocity equalization and active crosstalk cancellation. Fig. 2 exemplarily shows the transfer functions of the 20 IKO2 loudspeakers for one loudspeaker actively driven by an input signal.

[3] https://phaidra.kug.ac.at
[4] https://opendata.iem.at
[5] https://www.sofaconventions.org
[6] https://phaidra.kug.ac.at/detail_object/o:67609
[7] https://github.com/kronihias/mcfx
[8] https://github.com/kronihias/ambix

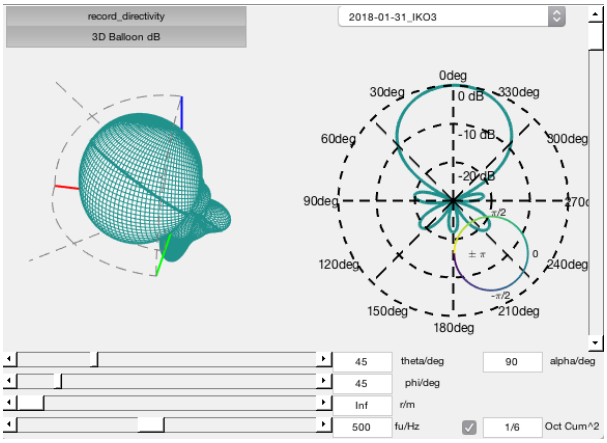


Fig. 1: Matlab based GUI of *balloon_holo* as analyzing tool for IKO beamforming.

2.2 Ambisonic Decoding Filters

A suitable filter design for a max-$\mathbf{r}_E$-weighted TOA decoding is discussed in [2], which uses modal gain limitation of the loudspeaker cone excursions (ranging from 0^{th} order for LF to 3^{rd} for HF). Instead of the trace of crosstalk matrix as in [2, p.63, right col], the filter design here uses a transfer function of a suitable electro-mechanical model as a -10 dB attenuated diagonal load for regularization. This is done to increase the robustness in the frequency range around 1.6 kHz, where the velocities exhibit pronounced notch whose full equalization might not always be useful, cf. Fig. 2. The filter design yields a 20x16 FIR matrix with 320 filter-and-sum operations of 4096 taps. The partitioned block convolution of the *mcfx_convolver* offers an efficient rendering thereof. Its efficiency relies on gathered

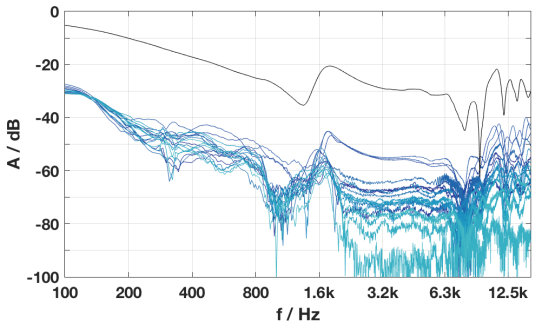


Fig. 2: Typical voltage to velocity transfer functions of IKO2 derived by laser vibrometry. Black...active loudspeaker, blue...passive loudspeakers.

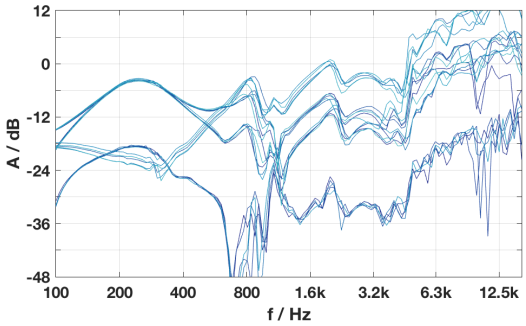


Fig. 3: Typical driving filters of the 20 IKO1 loudspeakers for beam steering to a vertex.

FFTs at the 16 inputs and IFFTs at the 20 outputs, instead of individual FFT-IFFT pairs for each of the 320 paths. Fig. 3 shows exemplary magnitude responses of the TOA decoding filters for a beam to an IKO1 vertex.

2.3 Sound Pressure Measurements

To analyze radiation characteristics, the IKOs were set up in a surrounding microphone array in a measurement chamber. Then directional impulse response from every loudspeaker to every microphone was measured by interleaved exponential sweeps. For IKO1 and IKO2, directional responses were obtained on an equiangular grid of 18x36 zenith and azimuth angles (648 points, spherical harmonics transform (SHT) order ≤ 17). For IKO3 the equiangular grid used 15x36 positions in zenith and azimuth (540 points, SHT order ≤ 14). Based on exterior SHT domain wave field extrapolation, the directional response is obtained in the far field. To involve the IKO's beam control, the input of the 20 directional far-field responses are convolved with the TOA decoding filters for a desired beam direction.

3 Results

The figures here are based on the following data sets:
IKO1: LV 12/2015, FIR 03/2018, MIC 08/2011
IKO2: LV 08/2017, FIR 03/2018, MIC 06/2016
IKO3: LV 01/2018, FIR 03/2018, MIC 01/2018
(LV...laser vibrometer measurement, FIR...TOA decoding filters, MIC...mic array measurement)

Directivity patterns in Fig. 4 for IKO1,2,3, and balloons for IKO3 in Fig. 5 were rendered using *balloon_holo*, cf. Fig. 1. Fig. 4 illustrates the slightly lower operational frequency range of the beams of IKO1 vs. those of IKO2,3 due to its 20% larger cabinet size. Main lobe differences of IKO3 to IKO2 due to its 1% larger cabinet size are negligible. Except for the expected kr-dependency, the velocity measurements, the filter transfer functions, and the resulting radiation patterns show similar characteristics.

Face beam denotes a beam directed on-axis with an IKO loudspeaker/facet, *edge beam* denotes a beam directed through the middle of an IKO edge, where two facets meet, while *vertex beam* denotes a beam directed through an IKO corner, where five facets meet. Patterns of these directions exhibit characteristic grating lobes as shown in the middle and bottom row of Fig. 5. A face beam produces prominent grating lobes at adjacent vertices. For edge and vertex beams prominent grating lobes occur in face direction. A vertex beam is somewhat super-directive due to the constructive interference of the five adjacent in-phase loudspeakers.

4 Summary

This contribution briefly discussed the efforts of collecting and analyzing open source data and software for the TOA beamformer IKO.

For research on (spherical) acoustic arrays at the IEM, the dedicated open data webpage `https://opendata.iem.at` linking to the repository `https://git.iem.at/groups/opendata` serves as reference to access all software, documentation, and Phaidra links to SOFA formated data.

References

[1] Wendt, F., Sharma, G. K., Frank, M., Zotter, F., and Höldrich, R., "Perception of Spatial Sound Phenomena Created by the Icosahedral Loudspeaker," *Computer Music Journal*, 41(1), pp. 76–88, 2017, https://doi:10.1162/COMJ_a_00396.

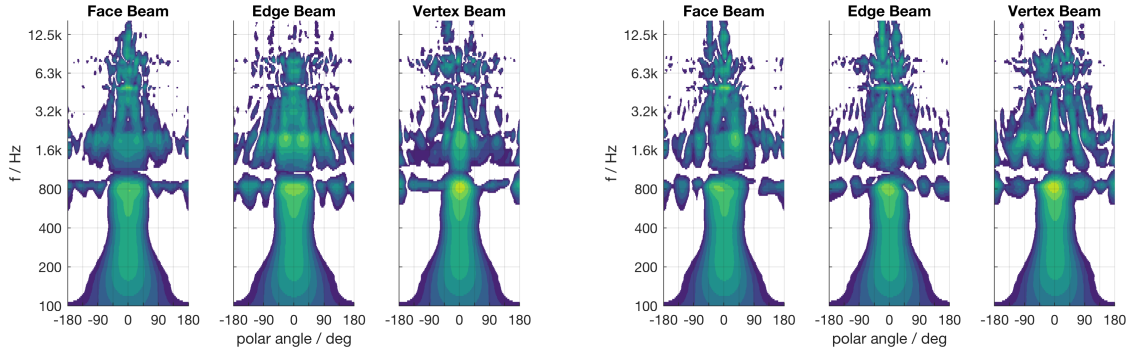
[2] Zotter, F., Zaunschirm, M., Frank, M., and Kronlachner, M., "A Beamformer to Play with Wall Reflections: The Icosahedral Loudspeaker," *Computer Music Journal*, 41(3), pp. 50–68, 2017, https://doi:10.1162/COMJ_a_00429.

[3] Wendt, F., Zotter, F., Frank, M., and Höldrich, R., "Auditory Distance Control Using a Variable-Directivity Loudspeaker," *Appl. Sci.*, 7(6), pp. 666–682, 2017, https://doi.org/10.3390/app7070666.

[4] Wendt, F., Zotter, F., Frank, M., and Höldrich, R., "Correction: Auditory Distance Control Using a Variable-Directivity Loudspeaker," *Appl. Sci.*, 7(11), p. 1174, 2017, https://doi.org/10.3390/app7111174.

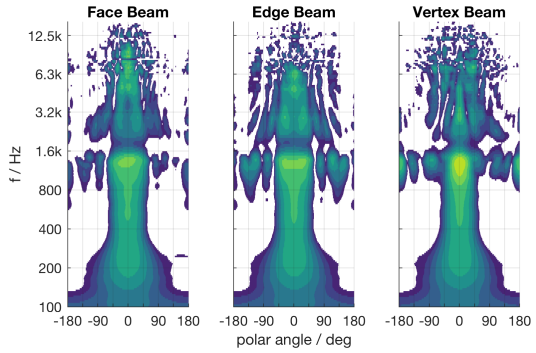
[5] Brandner, M., Frank, M., and Rudrich, D., "DirPat - Database and Viewer of 2D/3D Directivity Patterns of Sound Sources and Receivers," in *submitted E-Brief to Proc. of 144th Audio Eng. Soc. Conv. Milano*, 2018.

[6] Pollack, K., Meyer-Kahlen, N., and Zotter, F., "Design and Measurement of First-Order, Horizontally Beam-Controlling Loudspeaker Cube," in *submitted E-Brief to Proc. of 144th Audio Eng. Soc. Conv. Milano*, 2018.

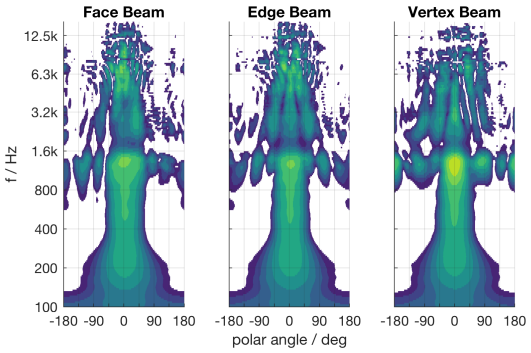


(a) IKO1 Horizontal Polar Slice.

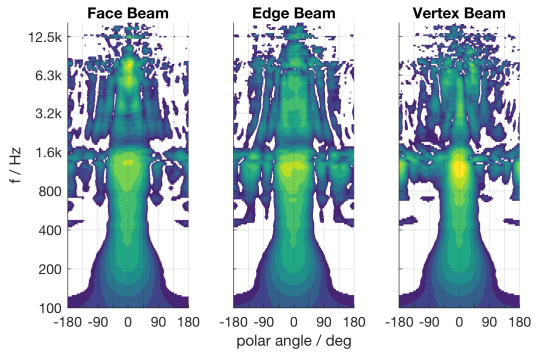
(b) IKO1 Vertical Polar Slice.



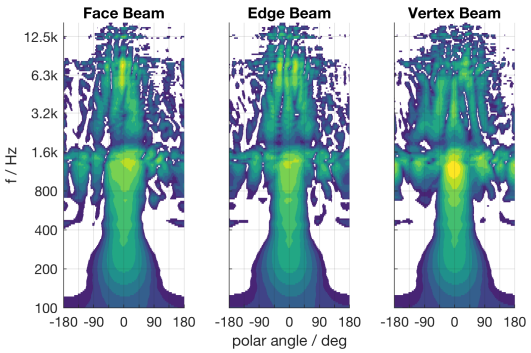
(c) IKO2 Horizontal Polar Slice.



(d) IKO2 Vertical Polar Slice.

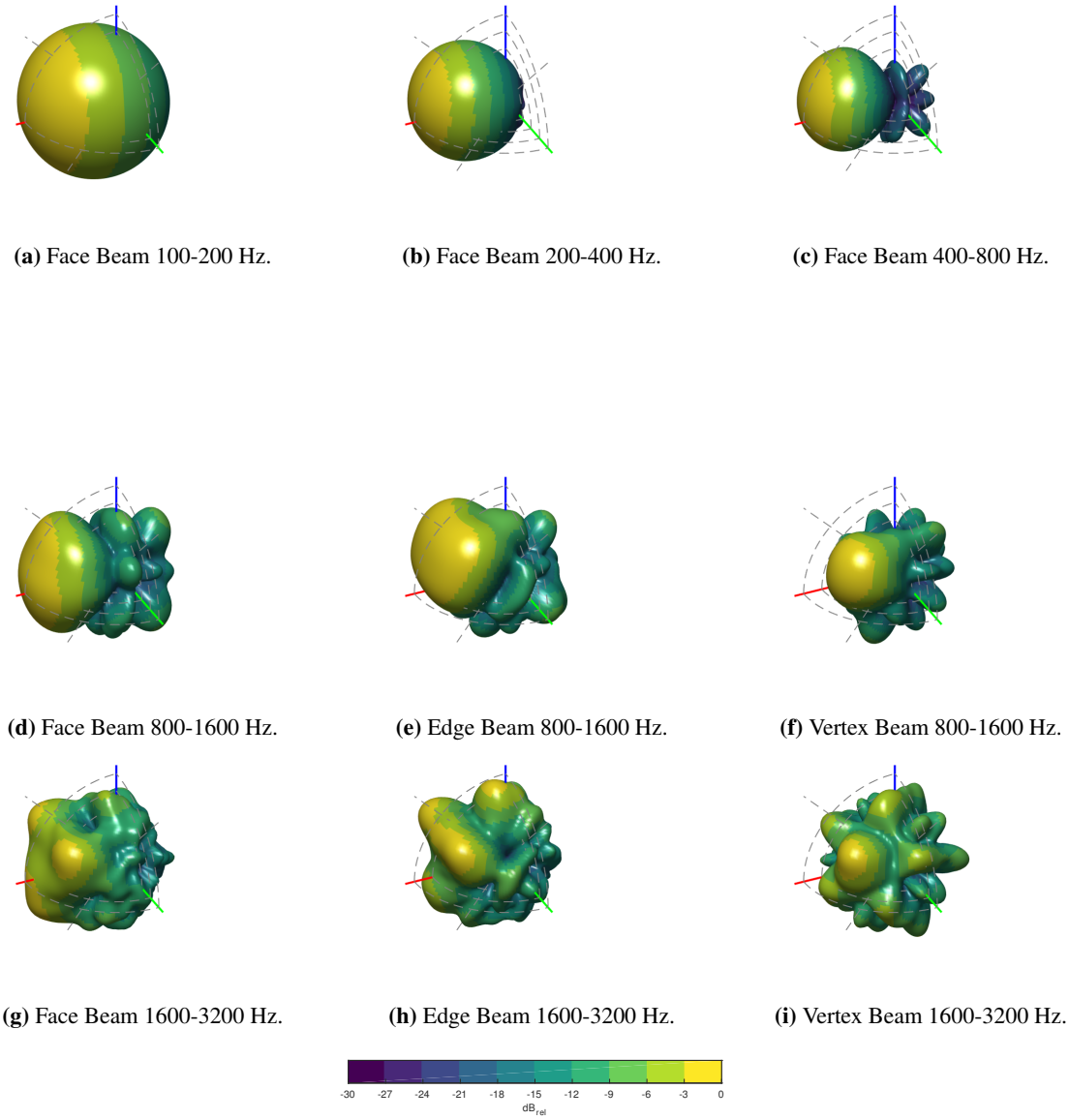


(e) IKO3 Horizontal Polar Slice.



(f) IKO3 Vertical Polar Slice.

Fig. 4: Horizontal and vertical directivity patterns for IKO1,2,3 using Ambisonic beamforming with frequency dependent order (cf. [2, Fig. 13]) and max $\mathbf{r}_E$-weighting. Colorbar indicates viridis colormap with 3 dB per color. Normalization for each subplot to a value derived from (i) energy in band 100 Hz to 400 Hz and (ii) averaging case (i) for $\pm 10^\circ$ along the main lobe. Overall normalization such that just no colormap clipping. Edge lengths: 0.345 m IKO1, 0.288 m IKO2, 0.294 m IKO3.



(a) Face Beam 100-200 Hz. **(b)** Face Beam 200-400 Hz. **(c)** Face Beam 400-800 Hz.

(d) Face Beam 800-1600 Hz. **(e)** Edge Beam 800-1600 Hz. **(f)** Vertex Beam 800-1600 Hz.

(g) Face Beam 1600-3200 Hz. **(h)** Edge Beam 1600-3200 Hz. **(i)** Vertex Beam 1600-3200 Hz.

Fig. 5: Directivity balloon characteristic for IKO3 using Ambisonic beamforming with frequency dependent order (cf. [2, Fig. 13]) and max $\mathbf{r}_E$-weighting. Colorbar indicates viridis colormap with 3 dB per color and 30 dB resolution. 3D polar diagram grid use 6 dB/div. The balloon radius and the color indicate dB-values of energy per frequency band defined by the given lower and upper cut frequency. Each subplot normalized to balloon's maximum dB value. Top row: up to about 1 kHz main lobe patterns are independent of the beam direction, exemplarily shown for a face beam here. Middle and bottom row: above 1 kHz frequency-dependent grating lobes arise that yield characteristic patterns for beams into face (left column), edge (center column) and vertex (right column) direction of the icosahedral shaped loudspeaker cabinet.

4

Binaural Rendering of Ambisonic Signals

4.1 Binaural Rendering of Ambisonic Signals by Head-Related Impulse Response Time Alignment and a Diffuseness Constraint

This work was published as:

M. Zaunschirm, C. Schörkhuber, and R. Höldrich (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.* doi:10.1121/1.5040489.

The idea and concept of this article were outlined by all authors. I, as first author wrote the original draft of the manuscript with help from the second author and periodical contributions from the third author. The second author contributed significantly to the third section of the article and formulated the solution of the constrained optimization problem. I did the programming, graphics, and objective evaluations.
The perceptual evaluation was outlined and discussed by all authors, while I did the programming and rendering, conducted the experiment, and did the statistical analysis. I wrote the revised version with periodic contributions of the second and third author.

Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint

Markus Zaunschirm, Christian Schörkhuber, and Robert Höldrich

Citation: The Journal of the Acoustical Society of America **143**, 3616 (2018); doi: 10.1121/1.5040489
View online: https://doi.org/10.1121/1.5040489
View Table of Contents: http://asa.scitation.org/toc/jas/143/6
Published by the Acoustical Society of America

Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint

Markus Zaunschirm,[a] Christian Schörkhuber, and Robert Höldrich
Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz, 8010, Austria

(Received 4 October 2017; revised 23 March 2018; accepted 17 May 2018; published online 19 June 2018)

Binaural rendering of Ambisonic signals is of great interest in the fields of virtual reality, immersive media, and virtual acoustics. Typically, the spatial order of head-related impulse responses (HRIRs) is considerably higher than the order of the Ambisonic signals. The resulting order reduction of the HRIRs has a detrimental effect on the binaurally rendered signals, and perceptual evaluations indicate limited externalization, localization accuracy, and altered timbre. In this contribution, a binaural renderer, which is computed using a frequency-dependent time alignment of HRIRs followed by a minimization of the squared error subject to a diffuse-field covariance matrix constraint, is presented. The frequency-dependent time alignment retains the interaural time difference (at low frequencies) and results in a HRIR set with lower spatial complexity, while the constrained optimization controls the diffuse-field behavior. Technical evaluations in terms of sound coloration, interaural level differences, diffuse-field response, and interaural coherence, as well as findings from formal listening experiments show a significant improvement of the proposed method compared to state-of-the-art methods. © *2018 Acoustical Society of America*.
https://doi.org/10.1121/1.5040489

[JB] Pages: 3616–3627

I. INTRODUCTION

Binaural rendering (synthesis) of acoustic scenes aims at evoking an immersive experience for listeners, which is desirable in virtual reality and 360-degree multimedia productions (Begault, 2000), and can also improve speech intelligibility in teleconferencing applications (Evans, 1997) by exploiting the effect of spatial unmasking (Freyman *et al.*, 1999; Litovsky, 2012).

Typically, binaural rendering involves a convolution of source signals with measured or modeled head-related impulse responses (HRIRs) or binaural room impulse responses (BRIRs) and playback via headphones (Møller, 1992; Wightman and Kistler, 1989). Both HRIRs and BRIRs implicitly contain the cues that are evaluated by the human auditory system to perceive sound from a certain direction and distance, with a certain source width, and spaciousness. A first theory on binaural perception was introduced by Rayleigh (1907) and is known today as the *Duplex theory*, which states that lateralization of sound sources is due to interaural time differences (ITD) for low frequencies, and due to the interaural level difference (ILD) at higher frequencies, although the frequency ranges of ITD and ILD localization overlap significantly. However, ILD and ITD information cannot be mapped directly to a certain direction, as sources that originate from the same *cone of confusion* evoke similar ILDs and ITDs, and thus, spectral cues are needed to resolve the ambiguities; a review is presented in Carlile *et al.* (2005). Besides the localization cues, the interaural coherence (IC) is related to the apparent source width, or parameters such as envelopment or spaciousness (Lindau, 2014; Okano *et al.*, 1998; Pollack and Trittipoe, 1959).

It has been previously shown that localization ambiguities can be reduced and externalization can be improved by using dynamic binaural rendering, where the natural head rotation of listeners is accounted for (Begault *et al.*, 2001; Brungart *et al.*, 2006; Wallach, 1940). Dynamic binaural rendering of object-based audio typically involves fading or switching of filters (HRIRs or BRIRs) (Engdegard *et al.*, 2008), while for dynamic binaural rendering of Ambisonic signals (scene-based audio), no filter switching is needed as the entire sound scene can be rotated by a simple frequency-independent matrix multiplication (Jot *et al.*, 1999; Pinchon and Hoggan, 2007).

Ambisonics (Daniel, 2000; Gerzon, 1973; Malham and Myatt, 1995) is based on the representation of a three-dimensional sound field with an orthonormal basis, the *spherical harmonics* (SH). Typically, the maximum SH order N used for representation determines the spatial resolution, the number of channels $(N+1)^2$, and the minimum number of loudspeakers required for playback. Binaural rendering of Ambisonics signals typically consists of decoding to virtual loudspeakers using a state-of-the-art method, e.g., Zotter and Frank (2012), followed by a summation of the virtual loudspeaker signals convolved with the HRIRs for the corresponding directions, see also Jot *et al.* (1998) and Noisternig *et al.* (2003). However, more recent methods (Bernschütz *et al.*, 2014; Sheaffer *et al.*, 2014) employ a rendering in the SH domain without the intermediate step of decoding to a virtual loudspeaker setup.

For direct rendering in the SH domain, the spatial order of the Ambisonic signals and HRIR description must match (Bernschütz *et al.*, 2014). It has been shown that HRIRs

[a]Electronic mail: zaunschirm@iem.at

represented in the SH domain contain a significant amount of energy in orders up to $N = 30$. However, recording, transmitting, and processing of 961 ($N = 30$) channels is usually not feasible. Thus, a low-order HRIR representation has to be used for rendering of practical Ambisonic orders $N < 5$. This low-order representation typically leads to impairments of localization cues (ILD, ITD), reduced spaciousness (IC), and a severe roll-off at high frequencies (Avni *et al.*, 2013; Bernschütz *et al.*, 2014; Sheaffer and Rafaely, 2014) (see also Sec. II A). Strategies which aim at improving the perceptual aspects of binaurally rendered order-limited Ambisonic signals include (i) a static equalization filter which is derived from the diffuse-field response of an HRIR set or spherical head model (Ben-Hur *et al.*, 2017; Sheaffer and Rafaely, 2014) and (ii) using a composite grid (spatial resampling) that matches the SH order N (Bernschütz *et al.*, 2014).

Further methods for reducing the SH order of the HRIR representation are known in the field of SH-based HRIR interpolation. The order reduction is achieved by an independent description of the minimum-phase response and linear-phase of the HRIRs; see Evans (1997), Evans *et al.* (1998), Jot *et al.* (1999), Romigh *et al.* (2015). In order to retain the ITD information in the rendered signals, the direction of sources must be known in advance, and thus these methods are well-suited for rendering of object-based audio or rendering of a parameterized sound scene (Laitinen and Pulkki, 2009; Pulkki, 2007), but not for direct binaural rendering of Ambisonic signals.

We suggest the computation of a binaural renderer for Ambisonic signals which consists of a HRIR preprocessing stage and an optimization stage. In the preprocessing stage, a frequency-dependent time alignment is applied to the original HRIRs; in the optimization stage, the binaural renderer is obtained by frequency domain minimization of the squared approximation error with respect to the high-frequency time-aligned HRIRs subject to quadratic constraints such that the diffuse-field properties of the rendered signals match the diffuse-field properties of a model. This article is structured as follows. In Sec. II, we address the challenges of binaural rendering of low-order Ambisonic signals and summarize state-of-the-art binaural rendering methods based on suggestions given in Ben-Hur *et al.* (2017), Bernschütz *et al.* (2014), and Sheaffer and Rafaely (2014). The proposed binaural renderer is presented in detail in Sec. III. The evaluation of the proposed renderer and a comparison to existing methods via technical measures is presented in Sec. IV. Finally, the findings from formal listening experiments are discussed in Sec. V.

II. BINAURAL RENDERING OF AMBISONIC SIGNALS

Let us consider a continuous distribution of sources in the far-field $s(\omega, \Omega)$, where ω is the frequency, $\Omega \equiv (\phi, \theta) \in \mathbb{S}^2$, $\phi = (0, 2\pi]$ is the azimuth angle, which is measured counter clockwise from the Cartesian x-axis in the horizontal plane, and $\theta = (\pi/2, -\pi/2)$ is the elevation angle, which increases upwards from the Cartesian x–y plane. With a continuous description of the far-field head-related transfer functions (HRTFs) $\boldsymbol{h}(\omega, \Omega) = [h^l(\Omega, \omega), h^r(\Omega, \omega)]^{\mathrm{T}}$, the ear signals are obtained by

$$\boldsymbol{x}(\omega) = \left[x^l(\omega), x^r(\omega)\right]^{\mathrm{T}} = \int_{\mathbb{S}^2} s(\omega, \Omega)\boldsymbol{h}(\omega, \Omega)\mathrm{d}\Omega, \tag{1}$$

where $\int_{\mathbb{S}^2}(\cdot)\mathrm{d}\Omega \equiv \int_0^{2\pi}\int_{\pi/2}^{-\pi/2}(\cdot)\cos\theta\mathrm{d}\theta\mathrm{d}\phi$, $(\cdot)^{l,r}$ indicates the left and right ear, respectively, and $(\cdot)^{\mathrm{T}}$ is the transpose operator. The corresponding acoustic scene in Ambisonics of order N_A is given by the $(N_A + 1)^2 \times 1$ vector

$$\boldsymbol{a}(\omega) = \int_{\mathbb{S}^2} s(\omega, \Omega)\boldsymbol{y}_{N_A}(\Omega)\mathrm{d}\Omega, \tag{2}$$

$$\boldsymbol{y}_{N_A}(\Omega) = \left[Y_0^0(\Omega), \ldots, Y_n^m(\Omega), \ldots, Y_{N_A}^{N_A}(\Omega)\right]^{\mathrm{T}}, \tag{3}$$

where $Y_n^m(\Omega)$ are the real-valued SHs (Williams, 1999)

$$Y_n^m(\Omega) = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}}P_n^{|m|}(\sin\theta) \times \begin{cases} \sin(|m|\phi) & m < 0, \\ \cos(m\phi) & m \geq 0, \end{cases} \tag{4}$$

where $P_n^m(\cdot)$ is the associated Legendre function, and $0 \leq n \leq N_A$, $-n \leq m \leq n$ are the order and degree, respectively.

As Ambisonics is a scene-based, and not an object-based format, the source signals and directions are typically not known (a parametric description of the scene can be obtained by the directional audio coding method from Laitinen and Pulkki, 2009, and Pulkki, 2007, but is beyond the scope of this article) and thus, a desired renderer is independent of the actual source signal $s(\omega, \Omega)$. The binaural rendering matrix $\boldsymbol{B}_{N_A}(\omega)$ yields the ear signals $\hat{\boldsymbol{x}}(\omega) = \boldsymbol{B}_{N_A}^{\mathrm{H}}(\omega)\boldsymbol{a}(\omega)$ from an Ambisonic signal of the order N_A. And it is obtained by solving a least-squares (LS) problem of the form

$$\boldsymbol{B}_{N_A}^{LS}(\omega) = \underset{\boldsymbol{B}_{N_A}(\omega)}{\arg\min}\int_{\mathbb{S}^2}||\boldsymbol{B}_{N_A}^{\mathrm{H}}(\omega)\boldsymbol{y}_{N_A}(\Omega) - \boldsymbol{h}(\omega, \Omega)||_{\mathrm{F}}^2\mathrm{d}\Omega, \tag{5}$$

where $||\cdot||_{\mathrm{F}}$ is the Frobenius norm. In most practical situations, only samples of the underlying continuous HRTFs are available and Eq. (5) is approximated by numerical integration

$$\boldsymbol{B}_{N_A}^{LS}(\omega) = \underset{\boldsymbol{B}_{N_A}(\omega)}{\arg\min}||(\boldsymbol{B}_{N_A}^{\mathrm{H}}(\omega)\boldsymbol{Y}_{N_A,P} - \boldsymbol{H}(\omega))\boldsymbol{W}^{1/2}||_{\mathrm{F}}^2, \tag{6}$$

where

$$\boldsymbol{H}(\omega) = \left[\boldsymbol{h}_1(\omega), \ldots, \boldsymbol{h}_p(\omega), \ldots, \boldsymbol{h}_P(\omega)\right], \tag{7}$$

$$\boldsymbol{h}_p(\omega) = \left[h^l(\Omega_p, \omega), h^r(\Omega_p, \omega)\right]^{\mathrm{T}} \tag{8}$$

is an arbitrary set of HRTFs that is defined at the discrete directions $\Omega_p \equiv (\phi_p, \theta_p) \in \mathbb{S}^2$ with $p = \{1, \ldots, P\}$ as the index of the available grid points,

$$\boldsymbol{Y}_{N_A,P} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p, \ldots, \boldsymbol{y}_P], \tag{9}$$

$$\boldsymbol{y}_p = \left[Y_0^0(\Omega_p), \ldots, Y_n^m(\Omega_p), \ldots, Y_{N_A}^{N_A}(\Omega_p)\right]^{\mathrm{T}}, \tag{10}$$

and $\boldsymbol{W}$ is a real-valued frequency-independent diagonal weighting matrix containing the quadrature weights. We note that selection of an optimal sampling grid for the sphere is an open research question as the sampling must approximate the integral well and also be practical (Duraiswami *et al.*, 2005; Fliege and Maier, 1999; Maday *et al.*, 2008; Zotkin *et al.*, 2009).

A. Binaural renderer

For the sake of readability the frequency dependency is not indicated in the remainder of this section. From Eq. (6), the optimal solution of the LS formulation is found as

$$\boldsymbol{B}_{N_A}^{LS} = (\boldsymbol{Y}_{N_A,P}\boldsymbol{W}\boldsymbol{Y}_{N_A,P}^{\mathrm{H}})^{-1}\boldsymbol{Y}_{N_A,P}\boldsymbol{W}\boldsymbol{H}^{\mathrm{H}}, \tag{11}$$

where $\boldsymbol{B}_{N_A}^{LS}$ contains the approximated SH expansion coefficients of the HRTFs (Ahrens *et al.*, 2012; Pollow *et al.*, 2012). Here we assume a high-resolution closed sampling grid $P > (N_A+1)^2$ that achieves $(\boldsymbol{Y}_{N_A}\boldsymbol{W}\boldsymbol{Y}_{N_A,P}^{\mathrm{H}}) \approx \boldsymbol{I}$ (orthogonality property of SH), where $\boldsymbol{I}$ is the identity matrix and thus, the LS solution can be further simplified to

$$\boldsymbol{B}_{N_A}^{LS} = \boldsymbol{Y}_{N_A,P}\boldsymbol{W}\boldsymbol{H}^{\mathrm{H}}, \tag{12}$$

and the binaurally rendered ear signals are

$$\hat{\boldsymbol{x}}_{N_A}^{LS} = (\boldsymbol{B}_{N_A}^{LS})^{\mathrm{H}}\boldsymbol{a}. \tag{13}$$

For a unity-gain plane wave impinging from direction Ω_q, the Ambisonic signals are defined as $\boldsymbol{a} = \boldsymbol{y}_q$, and by substituting Eq. (12) in Eq. (13), the rendered ear signals are

$$\hat{\boldsymbol{x}}_{N_A}^{LS} = H W \boldsymbol{Y}_{N_A,P}^{\mathrm{H}}\boldsymbol{y}_q = \boldsymbol{h}_{N_A,q}, \tag{14}$$

where $\boldsymbol{h}_{N_A,q}$ are the reconstructed HRTFs at direction Ω_q, which are obtained by a weighted summation of the original HRTFs. The directional weighting of that summation is defined by

$$\boldsymbol{g}_{N_A,q}^{P} = \boldsymbol{W}\boldsymbol{Y}_{N_A,P}^{\mathrm{H}}\boldsymbol{y}_q = \left[g_{N_A,q}^{(1)}, \ldots, g_{N_A,q}^{(p)}, \ldots, g_{N_A,q}^{(P)}\right]^{\mathrm{T}}. \tag{15}$$

Please note that Eq. (14) can be interpreted as an approximation method for reconstructing HRTFs corresponding to any direction. However, we do not intend to present a method for HRTF interpolation, but solely use the differences between the reconstructed and the original HRTFs as performance measures for binaural rendering of Ambisonic signals. For SH-based HRTF approximation and modeling, the readers are referred to Evans *et al.* (1998), Romigh (2012), and Zotkin *et al.* (2009).

Now let us consider a HRTF set (KU-100) which is measured at $P = 2702$ discrete directions arranged according to a Lebedev grid (Bernschütz, 2013). It can be seen in Fig. 1 that the SH order N_H in order to obtain near perfect binaural rendering increases as frequency increases, and that for the chosen HRTF set $N_H \approx 30$ is necessary (Bernschütz *et al.*, 2014). However, in practice, the maximum order of the HRTFs represented in the SH domain must match the order of the Ambisonic signals to be rendered for playback, cf. Eq. (13). Typically, the signals obtained from spherical microphone arrays are encoded in orders $N_A < 5$, and thus the reduction of order leads to severe loss of spatial detail, especially at higher frequencies. Moreover, a reduction of the order $N_A < N_H$ leads to a broader main-lobe of the directional weighting function and thus, more HRTF directions around the source direction Ω_q contribute substantially to the rendered ear signals. The directional weighting functions for orders $N_A = 1$ and $N_A = 5$ and $\Omega_q = (0°, 0°)$ are depicted in Fig. 2. Although the shape of the directional weighting is independent of the source-direction Ω_q, the effect on the rendered ear signals is strongly direction-dependent due to the off-center position of the ears. For a head radius $r_H = 8.5$ cm, we calculate the time offsets $\tau_p^{l,r}$ (time difference between the center and the ears) for each grid point p via a simple geometric model

$$\tau_p^r = \cos(\theta_p)\sin(\phi_p) r_H\, c^{-1}, \quad \tau_p^l = -\tau_p^r, \tag{16}$$

where $c = 343(\mathrm{m/s})$ is the speed of sound. Utilizing Eq. (15), the IR between all grid nodes P and an omnidirectional microphone at the position of the right ear is

$$x^r(t) = \sum_{p=1}^{P} g_{N_A,q}^{(p)}\, \delta(t - \tau_p^r), \tag{17}$$

where t is the time, and $\delta(\cdot)$ is the delta function. The obtained IRs, and corresponding transfer functions for directions $\Omega_q = (\phi_q, 0°)$ on the horizontal plane, and $N_A = 5$ are depicted in Figs. 3(a) and 3(b), respectively. For directions from the front and back, we observe strong colorations (low-pass behavior), as the time offsets for neighboring grid nodes which comprise the main-lobe (almost equally weighted) are highest. On the contrary, less coloration is expected for lateral directions as the variation of time offsets for nodes

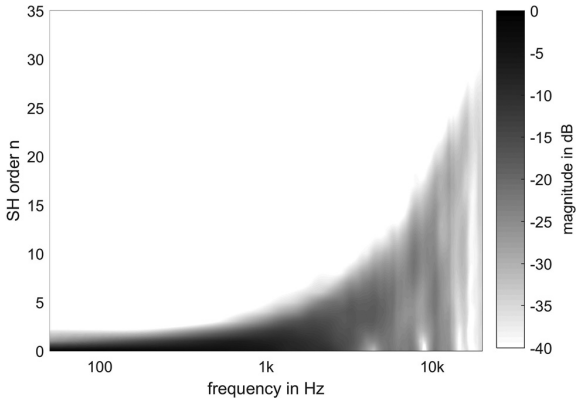


FIG. 1. Distribution of normalized energy contained in all modes $-n \le m \le n$ for each SH order n over frequency for the left ear and calculation of the SH expansion coefficients according to Eq. (11) for $N_A = 35$. ITD and head-shadowing effects (sharp dips and phase discontinuities at contralateral ear) lead to involvement of higher SH orders at high frequencies.

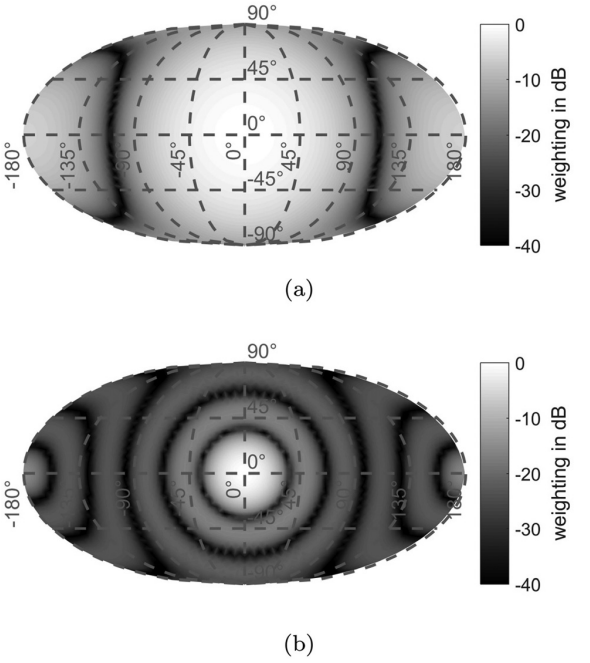


FIG. 2. Magnitude of the directional weighting function $\boldsymbol{g}_{N_A,q} = \boldsymbol{W}\boldsymbol{Y}_{N_A,P}^{\mathrm{H}}\boldsymbol{y}_q$ in dB for rendering of a plane wave impinging from $\Omega_q = (0°, 0°)$ using $N_A = 1$ (a), and $N_A = 5$ (b).

within the main-lobe is smaller. Similar findings are outlined in Solvang (2008), where an overview of the relation between number of loudspeakers, reproduction error, and coloration of two-dimensional Ambisonic systems is given.

Furthermore, the direction-dependent coloration for binaural rendering of low order Ambisonic signals is discussed in Avni *et al.* (2013), Ben-Hur *et al.* (2017), Bernschütz *et al.* (2014), and Sheaffer and Rafaely (2014) and improvements are obtained by (i) using a spectral (diffuse-field) equalization filter, which is based on a spherical head model (Ben-Hur *et al.*, 2017), or (ii) spatial resampling of HRTFs at a reduced grid (Bernschütz *et al.*, 2014). Both suggested methods are reviewed in the following paragraphs.

1. Spectral equalization

With the model of a perfectly diffuse sound field s_d consisting of an infinite number of plane waves impinging from every direction in space, with random mutually uncorrelated phases, and a total power ρ (see Epain and Jin, 2016), the covariance matrix of the ear signals $\boldsymbol{x}_d(t)$ is defined as $\boldsymbol{R}_H = E_t\{\boldsymbol{x}_d(t)\boldsymbol{x}_d^{\mathrm{H}}(t)\}$, where $E_t\{\cdot\}$ denotes the statistical expectation operator over the time t. With $\boldsymbol{x}_d(t) = \int_{\mathbb{S}^2} s_d(t, \Omega)\boldsymbol{h}(\Omega)d\Omega$, we get $\boldsymbol{R}_H = \rho\int_{\mathbb{S}^2}\boldsymbol{h}(\Omega)\boldsymbol{h}^{\mathrm{H}}(\Omega)d\Omega$, and by numerical integration and assuming $\rho = 1$, the estimated diffuse-field covariance matrix is obtained as

$$\boldsymbol{R}_H \cong \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}^{\mathrm{H}} = \begin{bmatrix} r^{ll}(\omega) & r^{lr}(\omega) \\ r^{rl}(\omega) & r^{rr}(\omega) \end{bmatrix}, \tag{18}$$

where the main diagonal entries contain the diffuse-field energies of the left and right ear, respectively. A similar formulation is used in Pulkki *et al.* (2017). With the diffuse-field

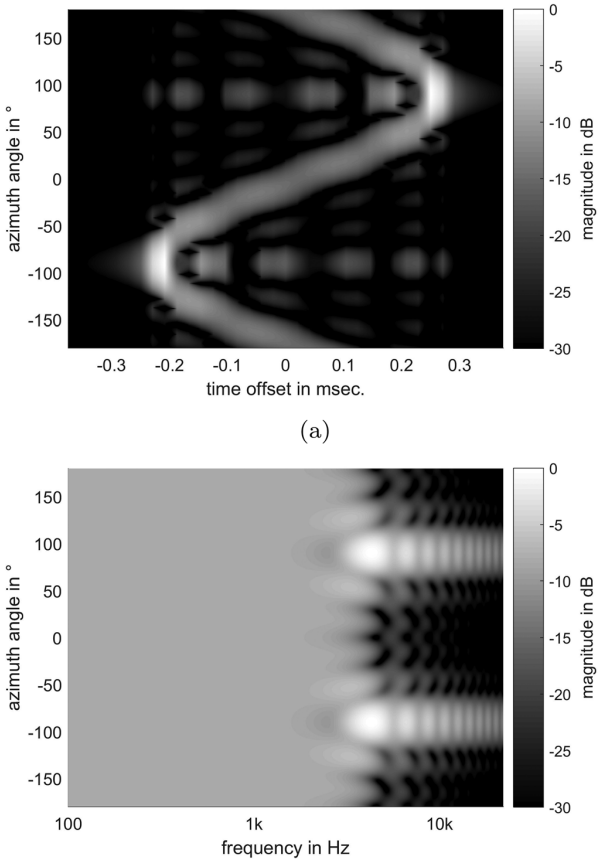


FIG. 3. (a) IRs between all grid nodes P and an omnidirectional microphone at the position of the right ear as defined in Eq. (17). The azimuth angle (ordinate) indicates the source-direction ϕ_q with $\theta_q = 0°$. The ideal IRs correspond to single pulses $\delta(t - \tau_p^r)$ with a time offset τ_p^r as defined in Eq. (16). (b) Transfer functions at the position of the right ear.

energies of the order N rendered signals $r_N^{qq}(\omega)$, where $q \in \{r, l\}$, the transfer function of the static diffuse-field equalization (timbre correction) filter (Sheaffer and Rafaely, 2014), which achieves $r_{N_A}^{qq}(\omega) = r_{N_H}^{qq}(\omega)$, where $N_A < N_H$ is given by

$$G^q(\omega)|_{N_A \to N_H} = \frac{\sqrt{r_{N_H}^{qq}(\omega)}}{\sqrt{r_{N_A}^{qq}(\omega)}}. \tag{19}$$

If instead of the measured HRTFs a rigid sphere head model is used, the diffuse-field energy according to Williams (1999) is

$$r_N^{qq}(\omega) = \sqrt{\sum_{n=0}^{N}(2n+1)|b_n(kr_H)|^2}, \tag{20}$$

where

$$b_n(kr_H) = 4\pi i^n\left[j_n(kr_H) - \frac{j_n'(kr_H)}{h_n'(kr_H)}h_n(kr_H)\right], \tag{21}$$

and $j_n(\cdot)$ is the spherical Bessel function, $h_n(\cdot)$ is the spherical Hankel function of the second kind, $j_n'(\cdot)$, $)h_n'(\cdot)$ are their first derivatives, and $k = \omega/c$.

In practice, we set $(G^l(\omega)|_{N_A\to N_H} + G^r(\omega)|_{N_A\to N_H})/2 = G(\omega)|_{N_A\to N_H}$ and the diffuse-field equalized (DEQ) binaural renderer is obtained by

$$\boldsymbol{B}_{N_A}^{DEQ} = G(\omega)|_{N_A\to N_H}\boldsymbol{B}_{N_A}^{LS}. \tag{22}$$

2. Spatial resampling

It has been found in Bernschütz *et al.* (2014) that selecting a feasible (sparse) set of grid points can reduce colorations. The HRTFs at the chosen composite grid nodes Ω_g, with $g = \{1,\ldots,G\}$, and $G < P$, are either selected from the original HRTFs (nearest neighbor), or are interpolated according to Eqs. (14) and (12) as

$$\boldsymbol{H}_{N_H,G} = (\boldsymbol{B}_{N_H}^{LS})^{\mathrm{H}}\boldsymbol{Y}_{N_H,G}. \tag{23}$$

In Bernschütz *et al.* (2014) an equiangular Gaussian quadrature (Stroud, 1966) with $G = 2(N_A+1)^2$ grid nodes and a Lebedev (Lebedev, 1977) quadrature were compared. In accordance with the results from listening experiments we use a composite Gaussian grid (CGG) binaural renderer, which is obtained by

$$\boldsymbol{B}_{N_A}^{CGG} = (\boldsymbol{Y}_{N_A,G}\boldsymbol{W}_G\boldsymbol{Y}_{N_A,G}^{\mathrm{H}})^{-1}\boldsymbol{Y}_{N_A,G}\boldsymbol{W}_G\boldsymbol{H}_{N_H,G}^{\mathrm{H}}, \tag{24}$$

where $\boldsymbol{W}_G$ contains the quadrature weights for the Gaussian sampling for the comparison of rendering methods.

III. PROPOSED BINAURAL RENDERER

The computation of the proposed binaural renderer consists of a preprocessing and an optimization stage. In the preprocessing stage, we apply a high-frequency time alignment to the original HRTF set. It has been shown in the context of SH-based HRTF interpolation that removing the linear phase (ITD equalization) leads to an order reduction of the HRTF representation. In contrast to Evans *et al.* (1998), Rasumow *et al.* (2014), and Romigh *et al.* (2015), we suggest a frequency-dependent ITD equalization that retains the ITD at low and removes ITD at high frequencies. Furthermore, the high frequency ITD is not re-synthesized after rendering.

Here, the time-aligned HRTF set is computed as

$$\mathring{h}l,r(\Omega_p,\omega) = h^{l,r}(\Omega_p,\omega)A_p^{l,r}(\omega), \tag{25}$$

where the frequency response of the allpass filter $A_p^{l,r}(\omega)$ is defined as

$$A_p^{l,r}(\omega) = \begin{cases} 1 & \text{for } \omega < \omega_c \\ e^{-i(\omega-\omega_c)\tau_p^{l,r}} & \text{for } \omega \geq \omega_c, \end{cases} \tag{26}$$

where $\omega_c = 2\pi f_c$, $i = \sqrt{-1}$, and the time offset $\tau_p^{l,r}$ is calculated according to Eq. (16). Note that the time offsets could be estimated from the HRTF set as well (see Katz and Noisternig, 2014 for a comparison of different methods). However, in pre-tests, we found no significant improvement and therefore used the simple geometric model. Due to the time alignment, the energy contained in higher SH orders is significantly reduced, cf. Figs. 1 and 4. Thus, lower orders $N_{\mathring{H}} < N_H$ are sufficient to represent the HRTFs at higher frequencies. As, according to the *Duplex Theory* (Hartmann *et al.*, 2016; Macpherson and Middlebrooks, 2002; Rayleigh, 1907; Wightman and Kistler, 1992), the ITD cue becomes less relevant as frequency increases, we expect that high-frequency time alignment of HRIRs with a cut-on frequency of $f_c = 1.5$ kHz (empirically chosen) allows for efficient order reduction while retaining the perceptually relevant localization cues.

In order to achieve the diffuse-field response and IC behavior of the original HRTF set $\boldsymbol{H}$ [see Eq. (18)], we cast the computation of the binaural renderer as a constrained optimization problem of the form

$$\boldsymbol{B}_{N_A}^{TAC} = \underset{\boldsymbol{B}_{N_A}}{\arg\min}\,||(\boldsymbol{B}_{N_A}^{\mathrm{H}}\boldsymbol{Y}_{N_A,P} - \mathring{\boldsymbol{H}})\boldsymbol{W}^{1/2}||_{\mathrm{F}}^2, \tag{27}$$

$$\text{subject to } \boldsymbol{B}_{N_A}^{\mathrm{H}}\boldsymbol{R}_Y\boldsymbol{B}_{N_A} = \boldsymbol{R}_H, \tag{28}$$

where $\mathring{\boldsymbol{H}}$ is the high-frequency time-aligned HRTF set, $\boldsymbol{R}_Y \equiv \boldsymbol{I}$ is the SH spatial covariance matrix, and $\boldsymbol{R}_H$ is defined in Eq. (18). A similar formulation is used in Schörkhuber and Höldrich (2017) and Vilkamo and Pulkki (2013). The set of solutions which satisfy the covariance constraint Eq. (28) is given by

$$\boldsymbol{B}_{N_A} = \boldsymbol{Q}\boldsymbol{C}, \tag{29}$$

where $\boldsymbol{Q}$ is an $(N_A+1)^2 \times 2$ arbitrary unitary matrix such that $\boldsymbol{Q}^{\mathrm{H}}\boldsymbol{Q} = \boldsymbol{I}$ and $\boldsymbol{C}$ is obtained by some suitable matrix decomposition of $\boldsymbol{R}_H$ such that $\boldsymbol{C}^{\mathrm{H}}\boldsymbol{C} = \boldsymbol{R}_H$. With the properties $||\boldsymbol{M}||_{\mathrm{F}}^2 = \mathrm{tr}\{\boldsymbol{M}^{\mathrm{H}}\boldsymbol{M}\}$ and $\mathrm{tr}\{\boldsymbol{M}_1+\boldsymbol{M}_2\} = \mathrm{tr}\{\boldsymbol{M}_1\} + \mathrm{tr}\{\boldsymbol{M}_2\}$, where $\mathrm{tr}\{\cdot\}$ is the trace of a matrix, and by inserting Eq. (29) into Eq. (27), we restate the minimization problem

$$\min_{\boldsymbol{Q}}\ \mathrm{tr}\{\boldsymbol{T}_1\} - 2\mathcal{R}(\mathrm{tr}\{\boldsymbol{T}_2\}) + \mathrm{tr}\{\boldsymbol{T}_3\}, \tag{30}$$

$$\text{s.t. } \boldsymbol{Q}^{\mathrm{H}}\boldsymbol{Q} = \boldsymbol{I}, \tag{31}$$

with

$$\boldsymbol{T}_1 = \boldsymbol{C}^{\mathrm{H}}\boldsymbol{Q}^{\mathrm{H}}\boldsymbol{Y}_{N_A,P}\boldsymbol{W}\boldsymbol{Y}_{N_H,P}^{\mathrm{H}}\boldsymbol{Q}\boldsymbol{C}, \tag{32}$$

$$\boldsymbol{T}_2 = \boldsymbol{C}\,\mathring{\boldsymbol{H}}\,\boldsymbol{W}\boldsymbol{Y}_{N_H,P}^{\mathrm{H}}\boldsymbol{Q}, \tag{33}$$

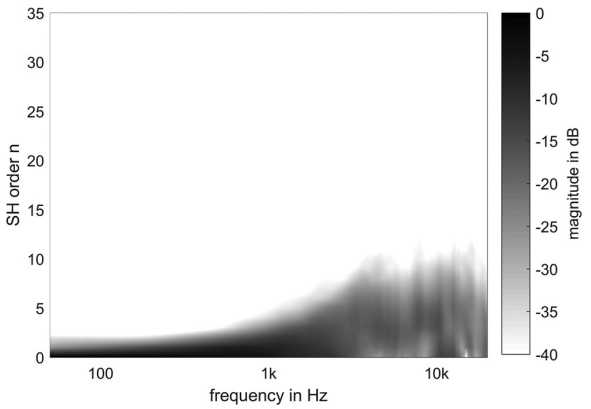


FIG. 4. Distribution of normalized energy contained in all modes $-n \leq m \leq n$ for each SH order n over frequency for the left ear and time-aligned HRIRs $\mathring{\boldsymbol{H}}$ (using $fc = 1.5$ kHz). The SH expansion coefficients are calculated according to Eq. (11). A lower order $N_{\mathring{H}} \approx 15$ is sufficient to represent the time-aligned HRIRs compared to the original set, cf. Fig. 1.

$$\boldsymbol{T}_3 = \mathring{\boldsymbol{H}}\boldsymbol{W}\mathring{\boldsymbol{H}}^{\mathrm{H}}, \tag{34}$$

where $\mathcal{R}(\cdot)$ denotes the real part of a complex number. Since $\boldsymbol{T}_3$ is independent of $\boldsymbol{Q}$, we drop it in the sequel. By assuming that $\boldsymbol{Y}_{N_A,P}\boldsymbol{W}\boldsymbol{Y}_{N_A,P}^{\mathrm{H}} = \boldsymbol{R}_Y = \boldsymbol{I}$, we can also drop the first term. Hence, the problem of determining $\boldsymbol{Q}$ is reduced to

$$\max_{\boldsymbol{Q}} \ \mathrm{tr}\{\boldsymbol{A}\boldsymbol{Q}\}, \tag{35}$$

$$\text{s.t. } \boldsymbol{Q}^{\mathrm{H}}\boldsymbol{Q} = \boldsymbol{I}, \tag{36}$$

where $\boldsymbol{A} = \boldsymbol{C}\mathring{\boldsymbol{H}}\boldsymbol{W}\boldsymbol{Y}_{N_A,P}^{\mathrm{H}}$. Using the singular value decomposition $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{H}} = \boldsymbol{A}$, the solution is given by

$$\boldsymbol{Q} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{H}}, \tag{37}$$

where $\boldsymbol{\Lambda} = [\boldsymbol{I}_2\, \boldsymbol{0}_{(N_A+1)^2-2\times 2}]^{\mathrm{T}}$. The final form of the time-aligned, and diffuse-field covariance-constrained (TAC) binaural renderer for order N_A is thus given by

$$\boldsymbol{B}_{N_A}^{TAC} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{H}}\boldsymbol{C}. \tag{38}$$

IV. EVALUATION

In this section, the rendered signals, which are obtained by the proposed TAC method are analyzed and compared with state-of-the-art methods presented in Sec. II. The quality criteria include (i) the direction-dependent coloration (presented for directions at the horizontal plane), (ii) the ILD errors in octave bands, and (iii) the diffuse-field behavior, i.e., the diffuse-field response and the interaural coherence.

A. Coloration

The composite loudness level (CLL) (Frank, 2013; Ono *et al.*, 2001, 2002) is a measure to describe the perceived timbre. We use the simplified definition

$$CLL_p(\omega) = |x^l(\Omega_p,\omega)|^2 + |x^r(\Omega_p,\omega)|^2, \tag{39}$$

where $x^{l,r}(\Omega_p,\omega)$ are the reference ear signals [see Eq. (1)] due to a single unity-gain plane wave impinging from direction Ω_p. The CLL error between the reference and rendered Ambisonic signals [see Eq. (13)] is defined as

$$ECLL_p^{N_A}(\omega) = 10\log\left(\frac{CLL_p^{N_A}(\omega)}{CLL_p(\omega)}\right), \tag{40}$$

where $CLL_p^{N_A}(\omega)$ is the CLL of the rendered signals using a rendering order N_A. The resulting CLL errors obtained for all discussed binaural rendering methods using $N_A = 3$ are depicted in Fig. 5 for directions on the horizontal plane ($\theta_p = 0°$). CLL errors for directions on the median plane show similar trends and are therefore not depicted here.

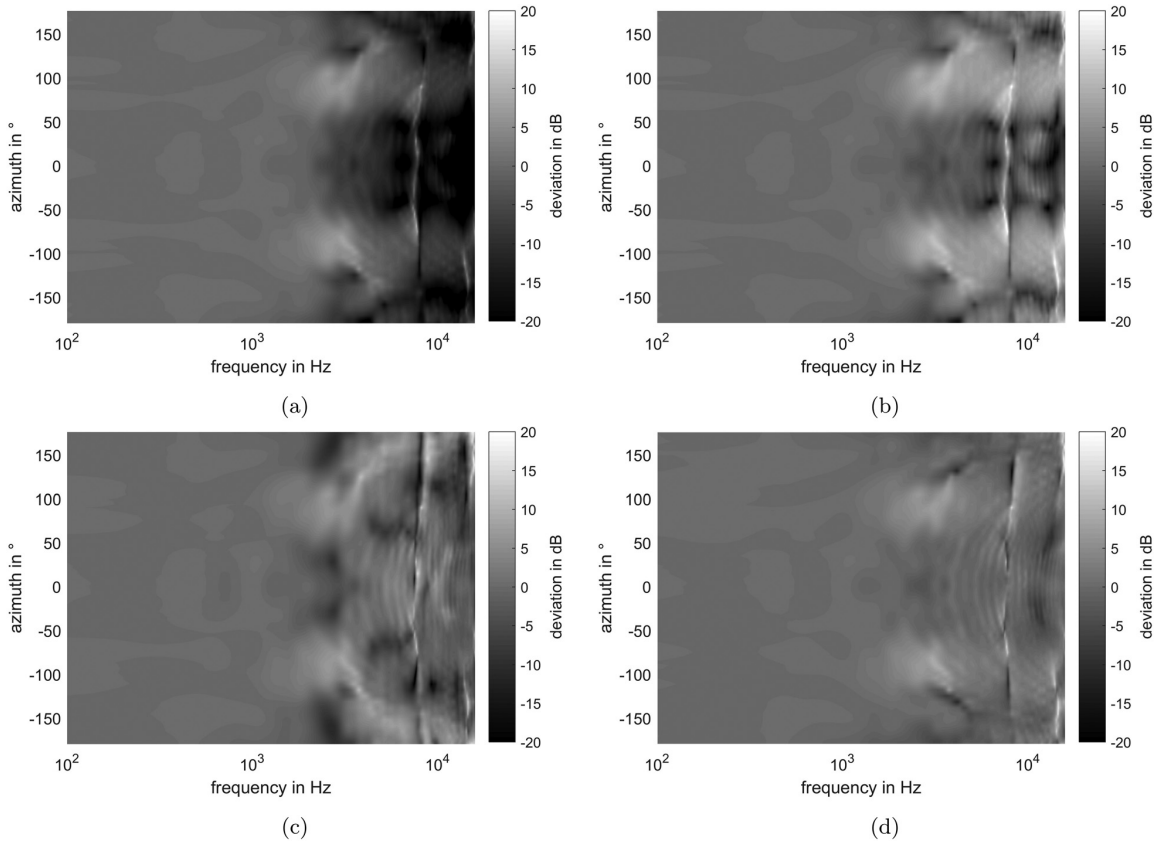
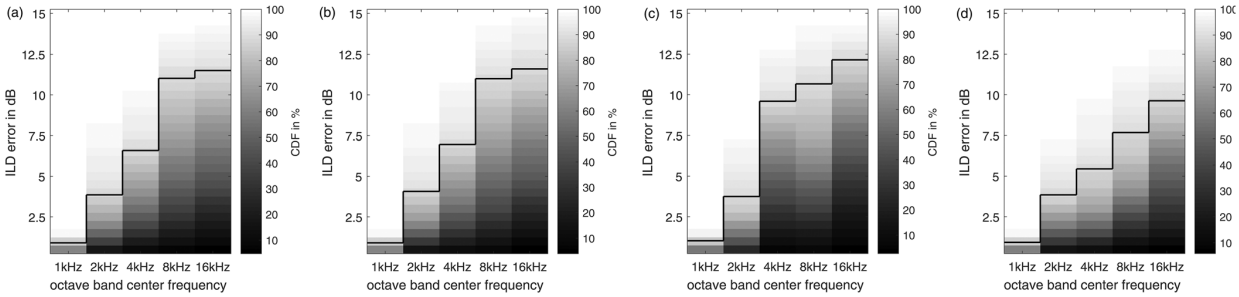


FIG. 5. CLL error according to Eq. (40) between reference and binaurally rendered Ambisonic signals in dB evaluated for sources at the horizontal plane ($\theta_q = 0°$) for an order of $N_A = 3$. (a) LS as defined in Eq. (11). (b) DEQ as defined in Eq. (22); see Ben-Hur *et al.* (2017). (c) CGG as defined in Eq. (24), see Bernschütz *et al.* (2014). (d) TAC as defined in Eq. (38).



(a) LS as defined in Eq. (11). (b) DEQ as defined in Eq. (22) (c) CGG as defined in Eq. (24) (d) TAC as defined in Eq. (38).

FIG. 6. Absolute ILD error between the original and the approximated HRIRs ($N_A = 3$) for all P directions is analyzed using a histogram between 0 and 15 dB using 30 equally spaced bins. Shown is the cumulative density function $v_i = \sum_{j=1}^{i}(c_j/P)$, where $P = 2702$ is the total number data points, c_j the number of elements in bin j, and i is the index of the histogram bins. The black solid lines indicate the 90% threshold of the CDF.

Below the aliasing frequency $f_{N_A} \approx N_A c/2\pi r_H \approx 1.9$ kHz (Rafaely, 2005) the CLL errors are negligible for all tested methods. Above the aliasing frequency, we observe a severe low-pass behavior for frontal and dorsal directions using the LS method, Eq. (11) [see Fig. 5(a)]. As the diffuse-field equalization [DEQ, see Eq. (22)] filter is basically a direction-independent high-shelving filter, the coloration error is shifted from frontal to lateral directions, cf. Fig. 5(b). The spatial resampling approach using a CGG as defined in Eq. (24) reduces the coloration for most directions, see Fig. 5(c). However, best performance in terms of minimal CLL error is observed for the proposed method (TAC), see Fig. 5(d).

B. ILD errors

The obtained ILD errors between the reference and rendered signals in octave-bands are defined as

$$EILD_p(\omega_o)^{N_A} = ILD_p(\omega_o) - ILD_p^{N_A}(\omega_o), \tag{41}$$

$$ILD_p(\omega_o) = 10\log\left(\frac{e_p^l(\omega_o)}{e_p^r(\omega_o)}\right), \tag{42}$$

where ω_o indicates an octave-band center frequency, and $e^{l,r}$ are the energies contained in the octave-bands of the left and right ear signal, respectively. The absolute values of the ILD errors are calculated for five octave-bands with center frequencies at 1, 2, 4, 8, and 16 kHz for all grid directions ($P = 2702$) and are analyzed with a histogram between 0 and 15 dB with 30 equally spaced bins. Figure 6 depicts the resulting cumulative density function (CDF) which is defined for each band as

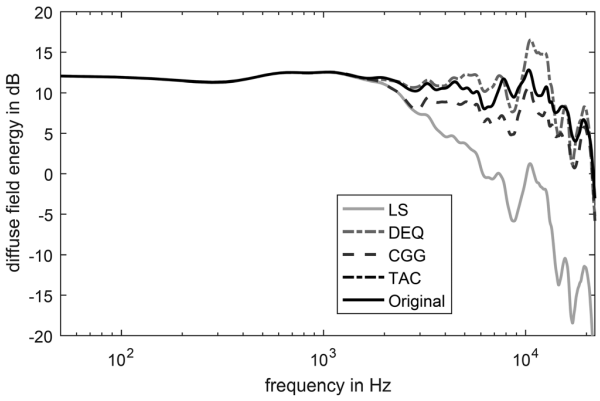
$$v_i = \sum_{j=1}^{i}\frac{c_j}{P}, \tag{43}$$

where c_j is the number of elements in bin j, and i is the histogram bin index.

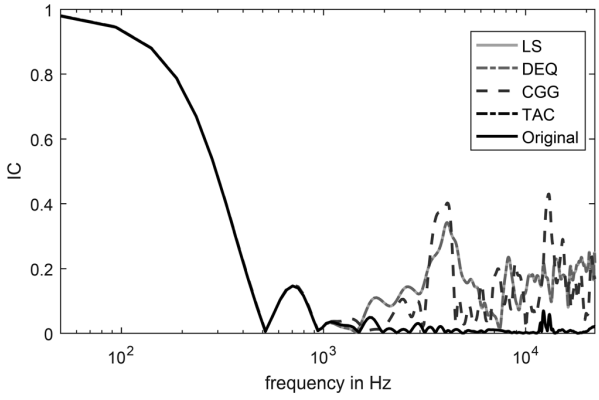
When comparing the sub figures depicted in Fig. 6, it can be observed that above the aliasing frequency $f_{N_A} \approx 1.9$ kHz ILD errors are increasing with frequency. Whereas for LS and DEQ, the distribution of absolute ILD errors is similar, rendering using the CGG approach shows the highest, and rendering using the TAC approach shows lowest overall ILD errors.

C. Diffuse-field response and interaural coherence

In order to compare the algorithms for rendering of diffuse sound fields, the main- and off-diagonal elements of the diffuse-field covariance matrix as defined in Eq. (18) are compared in Figs. 7(a) and 7(b), respectively. While the



(a)



(b)

FIG. 7. Diffuse-field energy (left ear) in dB (a) and interaural coherence (b) for a rendering order $N_A = 3$.

proposed TAC approach yields the same diffuse-field behavior as the reference set (due to constraint), all other approaches show deviations. The diffuse-field response of the LS renderer clearly indicates the discussed low-pass behavior. Results for DEQ and CGG show improvements, but above the aliasing frequency we observe colorations, cf. Fig. 7(a).

According to Menzer (2010) the interaural coherence (IC) is defined as

$$IC(\omega) = \frac{|r^{lr}(\omega)|}{\sqrt{r^{ll}(\omega) r^{rr}(\omega)}}, \tag{44}$$

where $r^{lr}(\omega)$, $r^{ll}(\omega)$, and $r^{rr}(\omega)$ are defined in Eq. (18). The IC of the tested rendering methods is depicted in Fig. 7(b). Again, the TAC yields the same behavior as the reference while the IC for LS and DEQ (same IC), and CGG shows significant deviations from the reference.

According to Gabriel and Colburn (1981), Pollack and Trittipoe (1959), and Stern *et al.* (2006), just noticeable differences (JNDs) of interaural correlation values change depending on the source frequency, bandwidth, and reference condition, and findings indicate a JND of 0.08 for a reference condition with interaural correlation of 1, and a JND of 0.35 for a reference condition with interaural correlation of 0, respectively. The JNDs for intermediate reference conditions are between 0.08 and 0.35 (Kim *et al.*, 2008).[1] As the IC deviations exceed the JNDs for certain frequency ranges, an altered spaciousness or envelopment is expected for LS, DEQ, and CGG rendering methods, especially for orders $N_A \leq 3$.

V. LISTENING EXPERIMENTS

In order to study and compare the perceptual aspects of binaural rendering using the TAC and state-of-the-art methods, formal listening experiments were conducted.

A. Methodology

Test participants were asked to rate the overall difference between a reference [rendered according to Eq. (11) with $N_A = N_H = 30$] and the test signals on a scale from *no audible difference* to *severe difference*. A hidden reference was used for screening of ratings, and thus the test procedure can be described as MUSHRA-like [multi stimulus with hidden reference and anchor (ITU-R, 1997)]. The presented test signals were continuously looped and participants were allowed to seamlessly switch between signals in real-time as often as desired.

The Ambisonic signals are obtained by a convolution of a monophonic source signal with a room impulse response (RIR) in the SH domain. For simulation of the RIRs, we used the multichannel room acoustics simulation toolbox (McRoomSIM) (Wabnitz *et al.*, 2010) for a shoebox room of dimensions $9.5 \times 12 \times 4.2$ m, with a mean absorption coefficient $\alpha = 0.2360$, a mean $T_{60} = 0.8$ s, and a source/listener setup as depicted in Fig. 8. The listener at position [3.5,3,1.7] m is facing towards the positive x-axis and the omnidirectional source is positioned relative to the listener as defined by the evaluation angle Ω_q on a radius $r_q = 1.5 r_c$, where $r_c = 1.39$ m is the critical distance. The tested discrete source directions include $\Omega_q = (0°, 0°)$, $\Omega_q = (90°, 0°)$, $\Omega_q = (35°, 45°)$, and $\Omega_q = (-45°, 0°)$. The perceptual evaluation is segmented in three experiments.

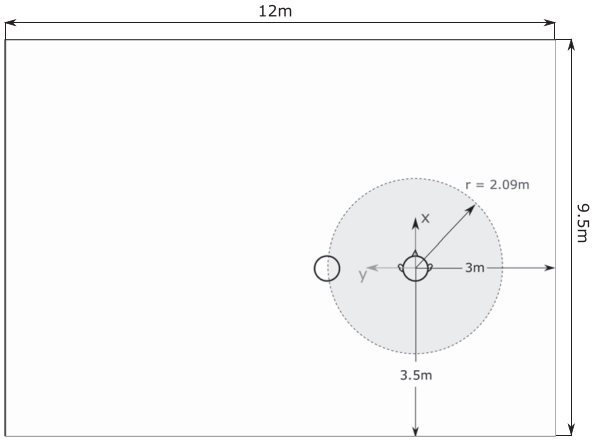


FIG. 8. Room layout and source/listener position used for simulating the room impulse responses via McRoomSIM. The listener is placed at [3.5,3,1.7]m and the source position is varied according to evaluation angle Ω_t on a radius $r_t = 1.5 r_c \approx 2.09$ m.

a. Experiment I. A speech signal and the direct part of the RIR were used for computing the test signals at all four test directions Ω_q. The tested methods were (i) LS, (ii) DEQ, (iii) CGG, and (iv) TAC for orders $N_A = 1$ and $N_A = 5$.

b. Experiment II. The test signals were a speech signal and a drum loop (kick drum, snare drum, cymbals). In order to evaluate the performance of the algorithms in reflective environments, the entire simulated RIR (direct, early reflections, and diffuse part) for all four test directions Ω_q was used. The tested algorithms include (i) DEQ, (ii) CGG, and (iii) TAC for orders $N_A = 1$, $N_A = 3$, and $N_A = 5$.

c. Experiment III. The dependence of the overall quality on the order $N_A = [1, 2, 3, 4, 5, 6, 9, 12, 15]$ was tested for the TAC method only. We used the entire RIR, the drum signal (as is it more complex), and $\Omega_q = (0°, 0°)$ and $\Omega_q = (90°, 0°)$ for testing.

Overall, 14 test pages were presented and the order of test signals within one page as well as the order of test pages were randomized. Depending on the experiment, the nine participants (expert listeners, no hearing impairments) were asked to rate the perceived overall difference on a continuous scale from no difference to severe difference for 9–10 test signals per page.

In order to ensure equal listening conditions for all participants and test signals, no head-tracking was used. This is valid as participants rated the difference to a reference and not the localization or externalization of stimuli. The test signals were played back via an AKG-702 (half-open) headphone and equalization according to Schärer and Lindau (2009) was used.

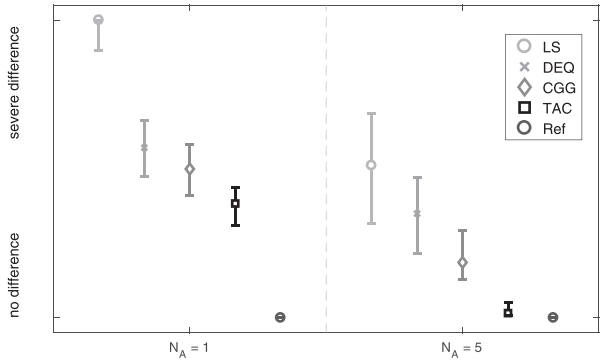


FIG. 9. Results of pooled data obtained in Experiment I showing the median (markers) and 95% confidence interval (solid lines) of ratings for all four tested directions using a speech signal and the direct-path only.

B. Results and discussion

The median and 95% confidence interval of all ratings (for all four test directions Ω_q) from Experiment I are depicted in Fig. 9. The results indicate a clear perceptual improvement for higher orders and that the proposed method (TAC) overall outperforms the other tested methods. The p-values (significance level) of a Kruskal-Wallis test (Kruskal and Wallis, 1952) presented in Table I indicate that there are five groups that are significantly different to each other.

The groups in ascending order of quality are (i) LS/1, (ii) DEQ/1, CGG/1, and LS/5, (iii) TAC/1, and DEQ/5, (iv) CGG/5, and (v) TAC/5. Overall, the TAC method yields the least perceptual differences to the reference for all tested orders and results for $N_A = 1$ are comparable to results for LS and DEQ using $N_A = 5$.

The detailed results (all test directions Ω_q separately) for Experiment I are shown in Fig. 10, where it can be seen that the performance of most methods varies with the source direction Ω_q. As most pronounced coloration of the LS approach is observed for frontal directions [see Fig. 10(a)] it is rated to have severe difference to the reference. On the other hand, the DEQ approach shifts the coloration from frontal to lateral directions and thus, performance is worst for $\Omega_q = (90°, 0°)$ [see Fig. 10(b)]. The CGG approach can give an improvement, but still the performance is highly dependent on the source direction [compare Figs. 10(a)–10(c)]. Results for the TAC method do not only show an overall improvement, but also the least variation across the different test directions.

TABLE I. p-values (Kruskal-Wallis) for all tested methods of Experiment I. The numbers after the / indicate the tested order N_A. Methods which are not significantly different (p-values > 0.05) are highlighted with gray rectangles.

LS/1	1.000							
LS/5	0.000	1.000						
DEQ/1	0.000	0.333	1.000					
DEQ/5	0.000	0.171	0.006	1.000				
CGG/1	0.000	0.800	0.123	0.096	1.000			
CGG/5	0.0000	0.001	0.000	0.042	0.000	1.000		
TAC/1	0.000	0.053	0.000	0.857	0.005	0.009	1.000	
TAC/5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
method	LS/1	LS/5	DEQ/1	DEQ/5	CGG/1	CGG/5	TAC/1	TAC/5

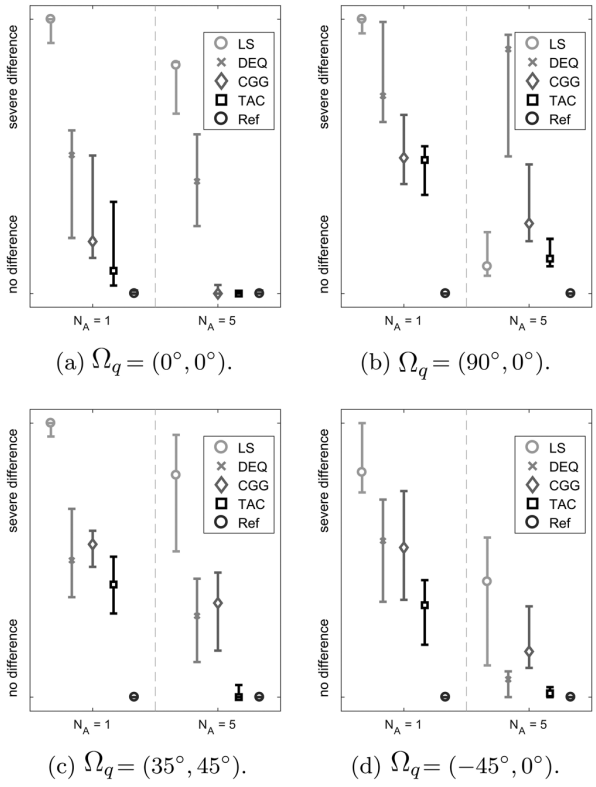


FIG. 10. Detailed results of Experiment I showing the median (markers) and 95% confidence interval (solid lines) of ratings from all participants for testing the perceived overall difference to the reference. The reference was a speech signal rendered for a far-field scenario and a source direction indicated by Ω_q.

The overall results for Experiment II are depicted in Fig. 11. Note that the results for the two different source signals (speech and drum loop), and all source directions Ω_q are pooled. We observe a similar behavior as for Experiment I, namely an improvement with increasing order N_A, and best

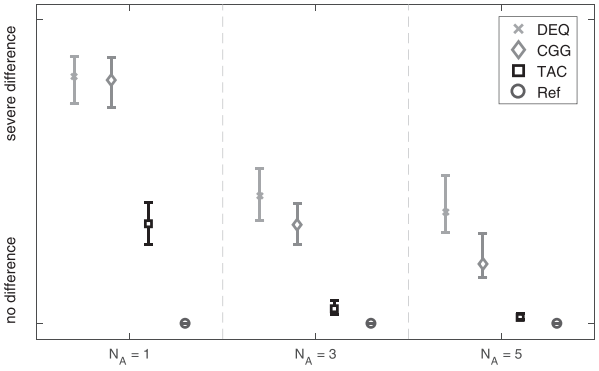


FIG. 11. Results of pooled data obtained in Experiment II showing the median (markers) and 95% confidence interval (solid lines) of ratings for all four tested directions using speech and drums as source signal and the entire simulated RIR.

TABLE II. p-values (Kruskal-Wallis) for all tested methods and pooled ratings of Experiment II. The numbers after the / indicate the tested order N_A. Methods which are not significantly different (p-values > 0.05) are highlighted with gray rectangles.

DEQ/1	1.000								
DEQ/3	0.000	1.000							
DEQ/5	0.000	0.994	1.000						
CGG/1	0.645	0.000	0.000	1.000					
CGG/3	0.000	0.084	0.138	0.000	1.000				
CGG/5	0.000	0.002	0.004	0.000	0.114	1.000			
TAC/1	0.000	0.084	0.143	0.000	0.952	0.105	1.000		
TAC/3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	
TAC/5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.008	1.000
method	DEQ/1	DEQ/3	DEQ/5	CGG/1	CGG/3	CGG/5	TAC/1	TAC/3	TAC/5

performance for the TAC method. The groups in ascending order of quality are (i) DEQ/1, and CGG/1, (ii) TAC/1, DEQ/3, CGG/3, DEQ/5, and CGG/5, (iii) TAC/3, and (iv) TAC/5. While ratings for the TAC method are significantly different across the tested orders $N_A = [1, 3, 5]$, there is no distinct difference between DEQ and CGG for orders $N_A = [3, 5]$; see Table II. Moreover, ratings for TAC and $N_A = 1$ are similar to DEQ and CGG for orders $N_A = 3$, and $N_A = 5$. The results per source signal are depicted in Fig. 12 and Fig. 13. Due to the transient and broadband nature of the drum signal (strong components above the aliasing frequency), the overall quality ratings are worse than for the speech signal. However, the TAC method shows smaller dependency on the source signal than the other tested methods.

The overall results for Experiment III are depicted in Fig. 14 for the two tested directions $\Omega_q = (0°, 0°)$ and $\Omega_q = (90°, 0°)$. As expected, the maximum order N_A to achieve near-transparent rendering changes with the source direction. The p-values listed in Table III indicate that for frontal sources, an order of $N_A = 4$ is sufficient (no significant difference to higher orders), but for lateral directions an order of $N_A = 9$ is required, see Fig. 14. As time-alignment of HRIRs reduces the required SH order for representation (see Fig. 4) to $N_{\mathring{H}} = 15$, testing of higher orders is not necessary.

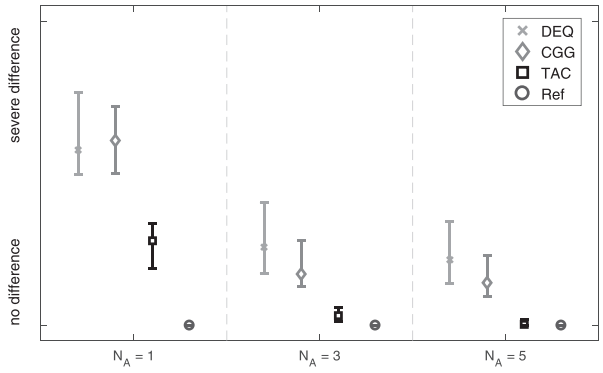


FIG. 12. Results of Experiment II showing the median (markers) and 95% confidence interval (solid lines) of ratings for all four tested directions using a speech signal and the entire RIR.

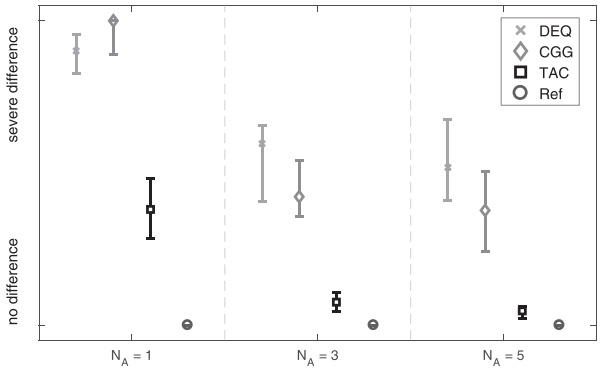


FIG. 13. Results of Experiment II showing the median (markers) and 95% confidence interval (solid lines) of ratings for all four tested directions using a drum signal and the entire RIR.

VI. CONCLUSION

In this paper, we presented an improved method for binaural rendering of low order Ambisonic signals ($N_A \leq 5$). The proposed binaural renderer is computed using a frequency-dependent time alignment of HRIRs followed by a minimization of the squared error subject to a diffuse-field covariance matrix constraint (TAC). Due to the time alignment, lower SH orders are sufficient to represent the directivity patterns of the ears at higher frequencies, while the covariance constraint ensures that sound scenes rendered with the TAC method achieve the same diffuse-field behavior as scenes rendered with the original high-order HRIRs.

Technical evaluations and comparisons to state-of-the-art methods indicate that the proposed TAC method reduces the direction-dependent colorations, and the ILD errors, and improves the diffuse-field behavior.

In the perceptual evaluation, we tested the overall difference to a reference (rendered with order $N_A = 30$) for four source directions in a free-field condition and in a simulated room. The results of the TAC method show a significant improvement of overall quality for all tested directions as well as smallest direction-dependent quality variation compared to the other tested methods. Furthermore, we found

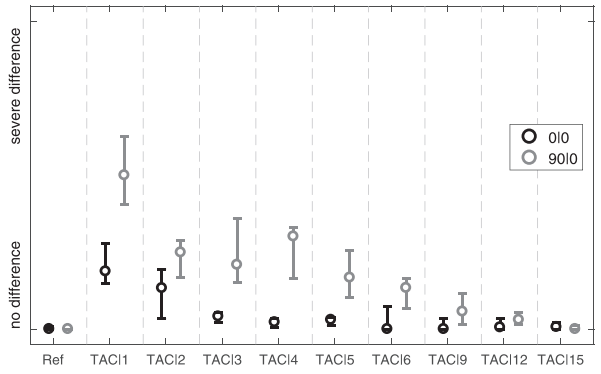


FIG. 14. Results of Experiment III showing the median (markers) and 95% confidence interval (solid lines) of ratings for all tested orders TAC|N_A using a drum loop as source signal and the entire RIR for rendering. Test directions $\Omega_q = (0°, 0°)$, and $\Omega_q = (90°, 0°)$ are depicted separately.

TABLE III. p-values (Kruskal-Wallis) for all tested orders of Experiment III. The upper triangle shows the p-values for $\Omega_q = (0°, 0°)$, the lower triangle for $\Omega_q = (90°, 0°)$ The numbers after the / indicate the tested order N_A. Methods which are not significantly different (p-values > 0.05) are highlighted with gray rectangles.

method	TAC/1	TAC/2	TAC/3	TAC/4	TAC/5	TAC/6	TAC/9	TAC/12	TAC/15
TAC/1	1.000	0.125	0.011	0.003	0.004	0.017	0.004	0.004	0.003
TAC/2	0.013	1.000	0.142	0.047	0.096	0.116	0.031	0.039	0.024
TAC/3	0.009	0.848	1.000	0.141	0.277	0.333	0.061	0.109	0.040
TAC/4	0.018	0.655	0.655	1.000	0.479	0.947	0.465	0.647	0.327
TAC/5	0.002	0.225	0.3062	0.142	1.000	0.647	0.267	0.245	0.174
TAC/6	0.002	0.025	0.073	0.035	0.482	1.000	0.620	0.946	0.838
TAC/9	0.002	0.013	0.018	0.013	0.125	0.179	1.000	0.733	0.838
TAC/12	0.002	0.002	0.003	0.003	0.006	0.009	0.305	1.000	0.894
TAC/15	0.002	0.002	0.002	0.002	0.004	0.004	0.089	0.077	1.000

that the rendering order N_A can be reduced significantly for the TAC method in order to achieve similar quality ratings as other binaural rendering methods for Ambisonic signals. Ratings of auralization in a simulated room indicate that the proposed method using $N_A = 1$ achieves comparable results to the other tested methods using $N_A = 5$.

As the TAC method shows little direction-dependent quality changes, we assume an improved externalization and localization performance. Thus, future work includes testing of the proposed method in a dynamic binaural rendering setup, where localization accuracy, externalization, and spaciousness are evaluated separately.

[1]The JNDs are defined for the single valued interaural correlation, and thus broadband signals. We assume that the frequency-dependent IC is an indicator for the interaural correlation for narrow-band signals.

Ahrens, J., Thomas, M. R., and Tashev, I. (**2012**). "HRTF magnitude modeling using a non-regularized least-squares fit of spherical harmonics coefficients on incomplete data," in *Proceedings of the Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, December 3–6, Hollywood, CA, pp. 1–5.

Avni, A., Ahrens, J., Geier, M., Spors, S., Wierstorf, H., and 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.

Begault, D. R. (**2000**). *3D Sound for Virtual Reality and Multimedia* (Academic Press, New York).

Begault, D. R., Wenzel, E. M., and Anderson, M. R. (**2001**). "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," J. Audio Eng. Soc. **49**(10), 904–916.

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

Bernschütz, B. (**2013**). "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," in *Proceedings the AIA-DAGA* 2013, March 18–21, Merano, Italy, pp. 592–595.

Bernschütz, B., Vazquez Giner, A., Pörschmann, C., and Arend, J. (**2014**). "Binaural reproduction of plane waves with reduced modal order," Acta Acust. united Acust. **100**(5), 972–983.

Brungart, D. S., Kordik, A. J., and Simpson, B. D. (**2006**). "Effects of head-tracker latency in virtual audio displays," J. Audio Eng. Soc. **54**(1–2), 32–44.

Carlile, S., Martin, R., and McAnally, K. (**2005**). "Spectral information in sound localization," Int. Rev. Neurobiol. **70**, 399–434.

Daniel, J. (**2000**). "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimedia" ("Representation of acoustic fields, application to the transmission and reproduction of complex soundscapes in a multimedia context"), Ph.D. thesis, University of Paris 6, Paris, France.

Duraiswami, R., Li, Z., and Zotkin, D. (**2005**). "Plane-wave decomposition analysis for spherical microphone arrays," Appl. Signal Process. Audio Acoust. **1**(5), 150–153.

Engdegard, J., Resch, B., Falch, C., Hellmuth, O., Hilpert, J., Hoelzer, A., Breebaart, J., Koppens, J., Schuijers, E., and Oomen, W. (**2008**). "Spatial audio object coding (SAOC): The upcoming MPEG standard on parametric object based audio coding," in *Proceedings of the 124th AES Convention*, May 17–20, Amsterdam, the Netherlands, pp. 1–15.

Epain, N., and Jin, C. T. (**2016**). "Spherical harmonic signal covariance and sound field diffuseness," IEEE Trans. Audio Speech Lang. Process. **24**(10), 1796–1807.

Evans, M. J. (**1997**). "The perceived performance of spatial audio for teleconferencing," Ph.D. thesis, University of York, York, UK.

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

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

Frank, M. (**2013**). "Phantom Sources using multiple loudspeakers in the horizontal plane," Ph.D. thesis, University of Music and Performing Arts, Graz, Austria.

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**(6), 3578–3588.

Gabriel, K. J., and Colburn, H. S. (**1981**). "Interaural correlation discrimination: 1. Bandwidth and level dependence," J. Acoust. Soc. Am. **69**, 1394–1401.

Gerzon, M. A. (**1973**). "Periphony: With-height sound reproduction," J. Audio Eng. Soc. **21**(1), 2–10.

Hartmann, W. M., Rakerd, B., Crawford, Z. D., and Zhang, P. X. (**2016**). "Transaural experiments and a revised duplex theory for the localization of low-frequency tones," J. Acoust. Soc. Am. **139**(2), 968–985.

ITU-R (**1997**). 1116-1: *Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems*, International Telecommunication Union, Geneva, Switzerland, pp. 1–26.

Jot, J.-M., Larcher, V., and Pernaux, J.-M. (**1999**). "A comparative study of 3-D audio encoding and rendering techniques," in *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction*, April 10–12, Arktikum, Finland, pp. 281–300.

Jot, J.-M., Wardle, S., and Larcher, V. (**1998**). "Approaches to binaural synthesis," in *Proceedings of the 105th Convention of the Audio Engineering Society*, September 26–29, San Francisco, CA, pp. 1–8.

Katz, B. F. G., and Noisternig, M. (**2014**). "A comparative study of interaural time delay estimation methods," J. Acoust. Soc. Am. **135**(6), 3530–3540.

Kim, C., Mason, R., and Brookes, T. (**2008**). "Initial investigation of signal capture techniques for objective measurement of spatial impression considering head movement," in *Proceedings of the 124th AES Convention*, May 17–20, Amsterdam, the Netherlands, pp. 1–17.

Kruskal, W. H., and Wallis, W. A. (**1952**). "Use of ranks in one-criterion variance analysis," J. Am. Stat. Assoc. **47**(260), 583–621.

Laitinen, M. V., and Pulkki, V. (**2009**). "Binaural reproduction for directional audio coding," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 18–21, New Platz, NY, pp. 337–340.

Lebedev, V. I. (**1977**). "Spherical quadrature formulas exact to orders 25–29," Sib. Math. J. **18**(1), 99–107.

Lindau, A. (**2014**). "Binaural resynthesis of acoustical environments—Technology and perceptual evaluation," Ph.D. thesis, University of Berlin, Berlin, Germany.

Litovsky, R. Y. (**2012**). "Spatial release from masking," Acoust. Today **8**(2), 18–25.

Macpherson, E. A., and Middlebrooks, J. C. (**2002**). "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," J. Acoust. Soc. Am. **111**(5), 2219–2236.

Maday, Y., Nguyen, N., Patera, A., and Pau, S. (**2008**). "A general multipurpose interpolation procedure: The magic points," Commun. Pure Appl. Anal. **8**(1), 383–404.

Malham, D. G., and Myatt, A. (**1995**). "3-D sound spatialization using Ambisonic techniques," Comput. Music J. **19**(4), 58–70.

Menzer, F. (**2010**). "Binaural audio signal processing using interaural coherence matching," Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.
Møller, H. (**1992**). "Fundamentals of binaural technology," Appl. Acoust. **36**(3–4), 171–218.
Noisternig, M., Musil, T., Sontacchi, A., and Höldrich, R. (**2003**). "3D binaural sound reproduction using a virtual ambisonic approach," in *Proceedings of the VECIMS 2003*, July 27–29, Lugano, Switzerland, pp. 174–178.
Okano, T., Beranek, L. L., and 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.
Ono, K., Pulkki, V., and Karjalainen, M. (**2001**). "Binaural modeling of multiple sound source perception: Methodology and coloration experiments," in *Proceedings of the AES 111th Convention*, November 30–December 3, New York, NY, pp. 1–12.
Ono, K., Pulkki, V., and Karjalainen, M. (**2002**). "Binaural modeling of multiple sound source perception: Coloration of wideband sound," in *Proceedings of the AES 112th Convention*, May 10–12, Munich, Germany, pp. 1–8.
Pinchon, D., and Hoggan, P. E. (**2007**). "Rotation matrices for real spherical harmonics: General rotations of atomic orbitals in space-fixed axes," J. Phys. A **40**(7), 1597–1610.
Pollack, I., and Trittipoe, W. (**1959**). "Binaural listening and interaural noise cross correlation," J. Acoust. Soc. Am. **31**(9), 1250–1252.
Pollow, M., Nguyen, K. V., Warusfel, O., Carpentier, T., Mueller-Trapet, M., Vorlaender, M., and Noisternig, M. (**2012**). "Calculation of head-related transfer functions for arbitrary field points using spherical harmonics decomposition," Acta Acust. united Acust. **98**(1), 72–82.
Pulkki, V. (**2007**). "Spatial sound reproduction with directional audio coding," J. Audio Eng. Soc. **55**(6), 503–516.
Pulkki, V., Delikaris-Manias, S., and Politis, A. (**2017**). *Parametric Time-Frequency Domain Spatial Audio* (John Wiley & Sons, New York).
Rafaely, B. (**2005**). "Analysis and design of spherical microphone arrays," IEEE Trans. Speech Audio Process. **13**(1), 135–143.
Rasumow, E., Blau, M., Hansen, M., van de Par, S., Doclo, S., Mellert, V., and Püschel, D. (**2014**). "Smoothing individual head-related transfer functions in the frequency and spatial domains," J. Acoust. Soc. Am. **135**(4), 2012–2025.
Rayleigh, L. (**1907**). "On our perception of sound direction," Philos. Mag. Ser. 6 **13**(74), 214–232.
Romigh, G. D. (**2012**). "Individualized head-related transfer functions: Efficient modeling and estimation from small sets of spatial samples," Ph.D. thesis, Carnegie Melon University, Pittsburg, PA.
Romigh, G. D., Brungart, D. S., Stern, R. M., and Simpson, B. D. (**2015**). "Efficient real spherical harmonic representation of head-related transfer functions," IEEE J. Selected Topics Signal Process. **9**(5), 921–930.
Schärer, Z., and Lindau, A. (**2009**). "Evaluation of equalization methods for binaural signals," in *Proceedings of the AES 126th Convention*, May 7–10, Munich, Germany, pp. 1–17.
Schörkhuber, C., and Höldrich, R. (**2017**). "Ambisonic microphone encoding with covariance constraint," in *Proceedings of the International Conference on Spatial Audio*, September 7–10, Graz, Austria, pp. 70–74.
Sheaffer, J., and Rafaely, B. (**2014**). "Equalization strategies for binaural room impulse response rendering using spherical arrays," in *Proceedings of the 28th Convention of Electrical and Electronics Engineers in Israel*, December 3–5, Eliat, Israel, pp. 1–5.
Sheaffer, J., Villeval, S., and Rafaely, B. (**2014**). "Rendering binaural room impulse responses from spherical microphone array recordings using timbre correction," in *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, April 3–5, Berlin, Germany, pp. 81–85.
Solvang, A. (**2008**). "Spectral impairment for two-dimensional higher order ambisonics," J. Audio Eng. Soc. **56**(4), 267–279.
Stern, R., Brown, G., and Wang, D. (**2006**). "Binaural sound localization," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications* (Wiley, New York), Chap. 5, pp. 147–185.
Stroud, A. H. (**1966**). *Gaussian Quadrature Formulas* (Prentice-Hall, Englewood Cliffs, NJ).
Vilkamo, J., and Pulkki, V. (**2013**). "Minimization of decorrelator artifacts in directional audio coding by covariance domain rendering," J. Audio Eng. Soc. **61**(9), 637–646.
Wabnitz, A., Epain, N., Jin, C. T., and Van Schaik, A. (**2010**). "Room acoustics simulation for multichannel microphone arrays," in *Proceedings of the International Symposium on Room Acoustics*, August 29–31, Melbourne, Australia, pp. 1–6.
Wallach, H. (**1940**). "The role of head movement and vestibular and visual cues in sound localization," J. Exp. Pyshcol. **27**(4), 339–368.
Wightman, F. L., and Kistler, D. J. (**1989**). "Headphone simulation of free field listening I: Stimulus synthesis," J. Acoust. Soc. Am. **85**, 858–867.
Wightman, F. L., and Kistler, D. J. (**1992**). "The dominant role of low frequency interaural time differences in sound localization," J. Acoust. Soc. Am. **91**(3), 1648–1661.
Williams, E. G. (**1999**). *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography* (Academic Press, New York).
Zotkin, D. N., Duraiswami, R., and Gumerov, N. A. (**2009**). "Regularized HRTF fitting using spherical harmonics," in *Proceedings of Applications of Signal Processing to Audio and Acoustics 2009*, October 18–21, New Platz, NY, pp. 257–260.
Zotter, F., and Frank, M. (**2012**). "All-round Ambisonic panning and decoding," J. Audio Eng. Soc. **60**(10), 807–820.

4.2 Binaural Rendering of Ambisonic Signals via Magnitude Least Squares

This work was published as:

C. Schörkhuber, **M. Zaunschirm**, and R. Höldrich (2018). Binaural Rendering of Ambisonic Signals via Magnitude Least Squares. *Proceedings of the DAGA*, 44(March):339–342.

The idea and concept of this article were outlined by the first author with contributions from the second and third author. I, as the second author, co-wrote the original draft and contributed significantly to the last section of the article (Optimal Cut-On Frequency). I programmed, conducted, and evaluated the listening experiment.

DAGA 2018 München

Binaural Rendering of Ambisonic Signals via Magnitude Least Squares

Christian Schörkhuber, Markus Zaunschirm, Robert Höldrich
Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz
schoerkhuber@iem.at

Introduction

Binaural rendering of order-limited Ambisonic signals is an active research area due to widespread adoption of the Ambisonic format for headphone-based reproduction of spatial audio content. When signal independent methods are considered, the problem is to find an optimal filter matrix which maps $J = (N+1)^2$ N-th order Ambisonic signals to the two ear signals. The two challenges in the underlying optimization process are (i) how to define a cost function that encodes perceptual dissimilarity, and (ii) how to find the global minimizer. In recent studies [1, 2, 3] it has been shown, that minimizing the squared error between the order N approximated head-related transfer functions (HRTFs) and a large set of measured/modeled head-related transfer functions (HRTFs) is a poor choice for low orders, as severe direction-dependent timbral artifacts are introduced. The observed rapid roll-off at high frequencies for frontal sources can be reduced by a global correction filter as suggested in [2], however, significant direction-dependent signal colorations remain. In [1] it was shown that signal colorations can be reduced by reducing the set of directions in the cost function. This approach is reminiscent of two-staged binaural rendering methods, where Ambisonic signals are first decoded to a set of virtual loudspeakers and then filtered with HFTFs corresponding to the loudspeaker directions. While these methods can rely on a rich body of research concerning loudspeaker-based reproduction of Ambisonic signals, it is unclear if these methods are optimal for headphone-based reproduction.
A rendering method specific to headphone-based reproduction is proposed in [3], where the colorations of rendered signals are significantly reduced by removing the linear phase from all HRTFs at higher frequencies prior to optimization. This approach is perceptually motivated as it can be assumed that altering the interaural phase difference (IPD) at higher frequencies is perceptually irrelevant [4]. Indeed, listening experiments conducted in [3] revealed, that the high-frequency phase modifications achieve a significant quality improvement over state-of-the-art methods.

The contribution of this work is twofold: Firstly, we evaluate the perceptual impact of high frequency phase modifications for different cut-on frequencies and signal types; and secondly, we show that the phase modification proposed in [3] can be viewed as an approximate solution of a noisy phase retrieval problem [5], and that the rendering quality can be further increased by solving the problem exactly.

Notation and Problem Formulation

In a nutshell, Ambisonic signals can be interpreted as the signals recorded by a set of virtual coincident microphones with directivity patterns that are proportional to spherical harmonics[1] up to some order $N \ll \infty$. That is, the j-th Ambisonic signal due to a plane wave signal $s_\Omega(\omega)$ from direction $\Omega = (\varphi, \theta)$, where φ, θ are the azimuth and elevation angle, respectively, is given by $a_j(\omega) = s_\Omega(\omega) Y_j(\Omega)$, where ω denotes frequency and $Y_j(\Omega) = Y_n^m(\Omega)$ is the spherical harmonic of order n and degree m evaluated at Ω, and $j = o(m,n)$ is a single index that depends on the ordering convention defined by the function $o(\cdot)$. In vector notation this can be written as $\boldsymbol{a}(\omega) = s_\Omega(\omega)\boldsymbol{y}(\Omega)$, where $\boldsymbol{a}(\omega) = [a_j(\omega)]_{j=1}^J$, $\boldsymbol{y}(\Omega) = [Y_j(\Omega)]_{j=1}^J$. With the target ear signals $b_l(\omega) = s_\Omega(\omega) H_l(\omega, \Omega)$ for $l \in \{\mathsf{L}, \mathsf{R}\}$, where $H_l(\omega, \Omega)$ is the measured/modeled HRTF, the goal is to find a rendering filter $\boldsymbol{w}(\omega)$ which yields an output signal $\hat{b}_l(\omega) = \boldsymbol{w}_l^{\mathsf{H}}(\omega)\boldsymbol{a}(\omega)$ that is *perceptually* as close as possible to the target signal $b_l(\omega)$. If the soundfield is modeled as a superposition of unknown plane wave signals from unknown directions, the problem can be written as

$$\boldsymbol{w}^*(\omega) = \arg\min_{\boldsymbol{w}} \int_{\Omega\in\mathcal{S}^2} D\left(\boldsymbol{w}^{\mathsf{H}}\boldsymbol{y}(\Omega),\ H(\omega,\Omega)\right) \mathrm{d}\Omega, \quad (1)$$

where $D(\cdot,\cdot)$ is a distance function which models the perceived dissimilarity. In the remainder we omit the dependency on ω for brevity, and we drop the subscript $l \in \{\mathsf{L}, \mathsf{R}\}$ as the HRTF set is assumed to be symmetric about the sagittal plane.

Least-Squares Methods

From a perceptual viewpoint it seems reasonable to define the distance function in (1) in terms of important binaural and monaural cues. However, defining and minimizing a perceptually motivated cost function is not trivial. Therefore, all methods proposed in [1, 2, 3] use some variation of a least-squares (LS) formulation, i.e.

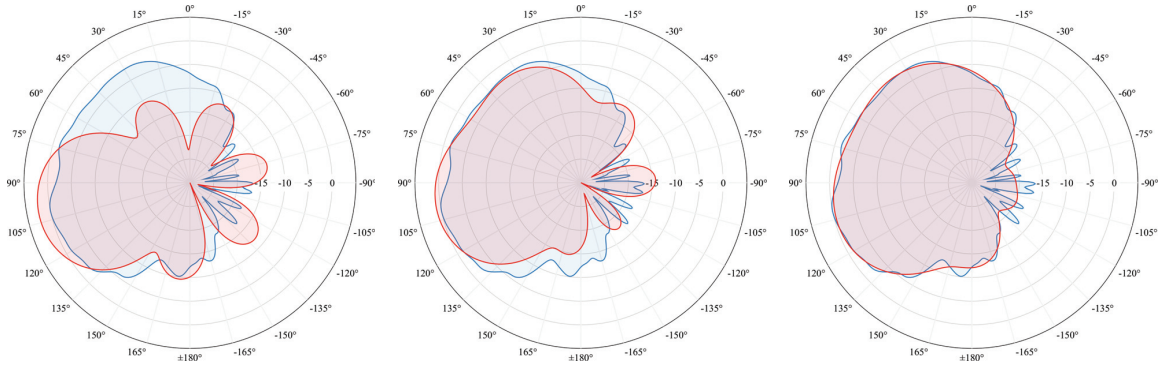
$$\min_{\boldsymbol{w}\in\mathcal{K}} \int_{\Omega\in\mathcal{S}^2} \left|\boldsymbol{w}^{\mathsf{H}}\boldsymbol{y}(\Omega) - H(\Omega)\right|^2 \mathrm{d}\Omega, \quad (2)$$

or its discrete approximation

$$\min_{\boldsymbol{w}\in\mathcal{K}} \sum_{\Omega\in\mathcal{M}} \left|\boldsymbol{w}^{\mathsf{H}}\boldsymbol{y}(\Omega) - H(\Omega)\right|^2 \quad (3)$$

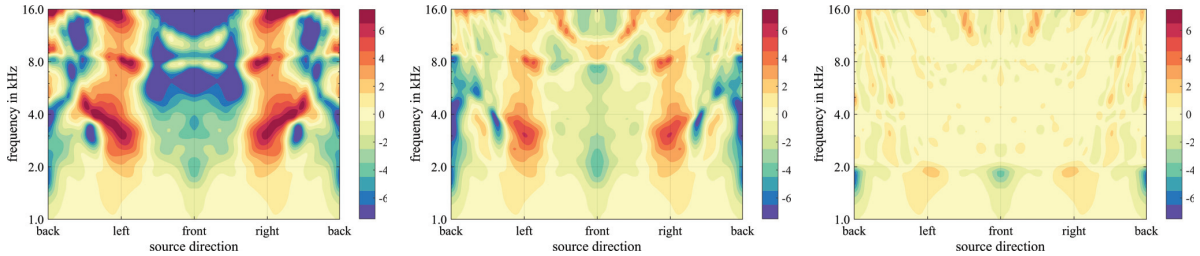
$$\equiv \min_{\boldsymbol{w}\in\mathcal{K}} \|\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w} - \boldsymbol{h}_{\mathcal{M}}\|_2^2, \quad (4)$$

[1] For ease of presentation, all numerical results in this contribution are given for the 2-dimensional case, i.e. we use circular harmonics rather than spherical harmonics.



(a) LS solution (equalized) [2] (b) LS solution (with time alignment) [3] (c) MLS solution (proposed)

Figure 1: Desired (blue) and approximated (red) HRTF magnitudes for $f = 6$ kHz and $N = 3$.



(a) LS solution (equalised) [2] (b) LS solution (with time alignment) [3] (c) MLS solution (proposed)

Figure 2: CLL error in dB for $N = 3$.

where $\mathcal{M}$ is a dense set of directions[2] such that $\boldsymbol{Y}_{\mathcal{M}}^{\mathsf{H}}\boldsymbol{Y}_{\mathcal{M}} = \boldsymbol{I}$, $\boldsymbol{Y}_{\mathcal{M}} = \left[\boldsymbol{y}^{\mathsf{H}}(\Omega)\right]_{\Omega\in\mathcal{M}} \in \mathbb{R}^{|\mathcal{M}|\times J}$, $\boldsymbol{h}_{\mathcal{M}} = \left[H(\Omega)\right]_{\Omega\in\mathcal{M}} \in \mathbb{C}^{|\mathcal{M}|}$, and $\mathcal{K}$ is the domain over which $\boldsymbol{w}$ is optimized. When the LS problem in (4) is unconstrained (i.e. $\mathcal{K} = \mathbb{C}^J$), the solution is given by

$$\boldsymbol{w}_{\mathsf{LS}} = \boldsymbol{Y}_{\mathcal{M}}^{\dagger}\boldsymbol{h}_{\mathcal{M}} = \boldsymbol{Y}_{\mathcal{M}}^{\mathsf{H}}\boldsymbol{h}_{\mathcal{M}} = \mathcal{SHT}_N^{\mathcal{M}}\left(\boldsymbol{h}_{\mathcal{M}}\right), \qquad (5)$$

where $(\cdot)^{\dagger}$ is a pseudo inverse, and $\mathcal{SHT}_N^{\mathcal{M}}(\cdot)$ denotes the order-N truncated discrete spherical harmonic transform (SHT). That is, the LS solution is equal to the SHT coefficients up to order N; alas, since the spatial complexity of HRTFs increases with frequency, a significant amount of energy is contained in modes up to order $N = 35$. Hence, the LS solution in (5) yields a severe spectral roll-off towards higher frequencies for low Ambisonic orders. To remedy this timbral artifact, the authors in [2] propose to apply a global diffuse-field equalization filter to the LS solution, which is equivalent to restricting the optimization domain in (4) to $\mathcal{K} = \left\{\boldsymbol{w}\in\mathbb{C}^J : \boldsymbol{w}^{\mathsf{H}}\boldsymbol{w} = \int_\Omega |H(\Omega)|^2 = \|\boldsymbol{h}_{\mathcal{M}}\|_2^2\right\}$ yielding the scaled (equalized) LS solution

$$\boldsymbol{w}_{\mathsf{LSeq}} = \frac{\|\boldsymbol{h}_{\mathcal{M}}\|_2}{\left\|\boldsymbol{Y}_{\mathcal{M}}^{\mathsf{H}}\boldsymbol{h}_{\mathcal{M}}\right\|_2}\boldsymbol{Y}_{\mathcal{M}}^{\mathsf{H}}\boldsymbol{h}_{\mathcal{M}}. \qquad (6)$$

The global equalization term in (6) reduces the overall spectral roll-off, but as the *local* modal order of HRTFs is direction-dependent - with higher orders for frontal directions due to rapid phase changes - direction-dependent colorations remain for lower Ambisonic orders.

The authors in [1] propose a different modification of the LS problem: rather than compensating for the loss of signal energy with a global filter, $\mathcal{M}$ in (4) is chosen to be sufficiently sparse such that higher-order modes are aliased down to lower-order modes. It has been shown, that this method effectively mitigates timbral artifacts, however, finding the optimal sparse set $\mathcal{M}$ (number of sampling points, sampling scheme, rotation) is not straight-forward as it influences both monaural and binaural cues.

Recently, in [3] another variation of the LS problem has been proposed that yields a significant improvement of the perceived quality. The basic notion is to modify the target HRTF set $H(\Omega)$ prior to optimization such that the energy in higher orders is reduced with minimal perceptual ramifications. The proposed HRTF modifications are based on two observations:

- Most of the energy in higher order modes is caused by rapid phase changes towards higher frequencies due to the off center location of the ears.
- With increasing frequency, the perceptual importance of interaural time differences (ITDs) decreases, while the relative importance of interaural level differences (ILDs) increases.

Consequently, a two-band HRTF modification scheme, referred to as time alignment (TA), has been proposed in

[2] We assume that the set $\mathcal{M}$ is chosen such that the integral is trivially approximated, e.g. by using a spherical t-design with high order. In practice, however, quadrature weights need to be used and the accuracy of the approximation depends on the number of sampling points relative to the modal order of the integrand.

[3], where the modified HRTFs are defined as

$$\tilde{H}(\omega,\Omega) = \begin{cases} H(\omega,\Omega) & \text{if } \omega \leq \omega_c \\ H(\omega,\Omega)e^{-i\phi_l(\omega,\Omega)} & \text{if } \omega > \omega_c, \end{cases} \quad (7)$$

where $\phi_l(\omega,\Omega)$ is the linear phase due to the ITD corresponding to direction Ω, and ω_c is the cut-on frequency above which phase modification is applied. By subtracting the linear phase part for high frequencies only, the important ITD cues are preserved at low frequencies while signal energy in higher-order modes is significantly reduced. This in turn reduces the energy loss when the modal order is truncated, thus reducing the spectral roll-off towards higher frequencies.

Proposed Method

We can write the cost function in (4) as

$$\min_{\boldsymbol{w}\in\mathcal{K}} \|\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w} - \boldsymbol{M}\boldsymbol{p}\|_2^2, \quad (8)$$

where $\boldsymbol{M} = \mathrm{diag}\left(|\boldsymbol{h}_{\mathcal{M}}|\right)$, $\boldsymbol{p} = \left(e^{i\gamma(\Omega)}\right)_{\Omega\in\mathcal{M}}$, and usually $\gamma(\Omega) = \angle H(\Omega)$. However, it has been shown in [3], that defining $\gamma(\Omega) = \angle H(\Omega) - \phi_l(\Omega)$ at higher frequencies significantly improves the quality of the solution. While removing the linear phase part from the HRTFs to reduce the modal order is conceptually well motivated, it is not clear which HRTFs phases yield the lowest spatial complexity. Therefore we propose to reformulate the LS problem as a joint minimization over $\boldsymbol{w}$ and $\boldsymbol{p}$, i.e.

$$\min_{\boldsymbol{w},\boldsymbol{p}} \|\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w} - \boldsymbol{M}\boldsymbol{p}\|_2^2 \quad (9)$$

$$\text{s.t. } |p_j| = 1 \;\; \forall j = 1,\ldots,|\mathcal{M}|, \quad (10)$$

where we simultaneously seek the optimal HRTF phase modification and the corresponding optimal filter coefficients. This quadratically constrained quadratic program (QCQP) might be solved via the popular semidefinite relaxation method [6], however, by observing that this problem is equivalent to the phase-lift formulation of the noisy phase retrieval problem [5], we can rewrite (9)-(10) as

$$\min_{\boldsymbol{w}} \| \, |\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w}| - |\boldsymbol{h}_{\mathcal{M}}| \, \|_2^2, \quad (11)$$

where the objective is to approximate only HRTF magnitudes while ignoring the phase error. The proposed rendering filters, referred to as magnitude least squares (MLS) solution, are thus given by

$$\boldsymbol{w}_{\mathsf{MLS}}(\omega_k) = \arg\min_{\boldsymbol{w}} \Big[\lambda(\omega_k) \, \| \, \boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w} - \boldsymbol{h}_{\mathcal{M}} \, \|_2^2 + (1-\lambda(\omega_k)) \, \| \, |\boldsymbol{Y}_{\mathcal{M}}\boldsymbol{w}| - |\boldsymbol{h}_{\mathcal{M}}| \, \|_2^2\Big], \quad (12)$$

where ω_k is the center frequency of the k-th bin, and $\lambda(\omega) = 1$ if $\omega \leq \omega_c$ and 0 otherwise.[3] Finding the global optimizer of (12) for $\lambda(\omega) < 1$ is non-trivial in general, however, we found that any local nonlinear optimization method can be used when the initial estimate for $\boldsymbol{w}_{\mathsf{MLS}}(\omega_k)$ is set to $\boldsymbol{w}_{\mathsf{MLS}}(\omega_{k-1})$. Note that the cost function in (12) is invariant under a global phase change if $\lambda(\omega_k) = 0$, and thus, we apply a phase rotation to the initial solution in order to obtain a smooth phase evolution:

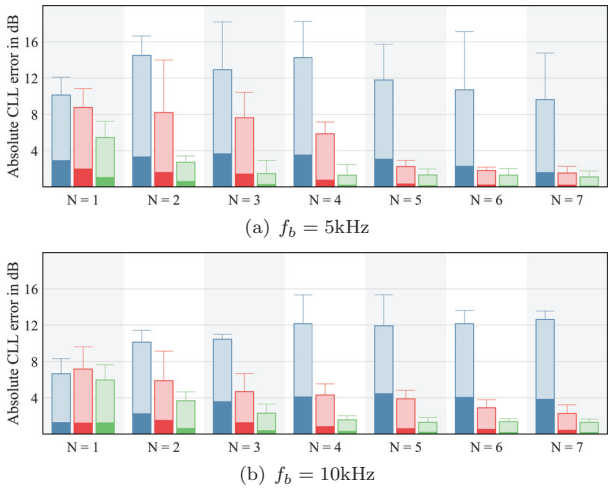
$$\boldsymbol{w}_{\mathsf{MLS}}(\omega_k) \leftarrow \boldsymbol{w}_{\mathsf{MLS}}(\omega_k)e^{i\zeta}, \quad (13)$$

where

$$\zeta = \arg\min_{\tilde{\zeta}} \left\| \boldsymbol{w}_{\mathsf{MLS}}(\omega_k)e^{i\tilde{\zeta}} - \boldsymbol{w}_{\mathsf{MLS}}(\omega_{k-1}) \right\|_2^2 \quad (14)$$

$$= \angle\left(\boldsymbol{w}_{\mathsf{MLS}}(\omega_k)^{\mathsf{H}}\boldsymbol{w}_{\mathsf{MLS}}(\omega_{k-1})\right). \quad (15)$$

[3] To avoid rapid filter changes around ω_c we choose a smooth transistion function in practice.



(a) $f_b = 5$kHz

(b) $f_b = 10$kHz

Figure 3: Statistics of the absolute CLL error for all directions and frequencies within octave band around f_b (median, 99th percentile, max). Blue: LS solution (equalized) | Red: LS solution (time aligned) | Green: MLS solution (proposed).

Numerical Evaluation

In Fig. 1 the magnitudes of the approximated HRTFs $\hat{H}(\Omega) = \boldsymbol{w}^{\mathsf{H}}\boldsymbol{y}(\Omega)$ are compared with the magnitudes of the target HRTFs $H(\Omega)$ for $N = 3$ and $\omega_k = 6$ kHz. It can be observed that subtracting the linear phase from the HRTFs significantly improves the approximation (Fig. 1(b)) compared to the LS solution with diffuse-field equalization in Fig. 1(a), and that the proposed MLS approach (Fig. 1(c)) reduces the magnitude error even further. In order to evaluate timbral artifacts for different source directions and frequencies, we use the composite loudness level (CLL), defined as

$$\mathrm{CLL}\left(H(\omega,\Omega)\right) = 10\log\left(|H(\omega,\Omega)|^2 + |H(\omega,\Omega')|^2\right),$$

where $\Omega' = (-\varphi,\theta)$, which is related to the perceived timbre. In Fig. 2 the CLL error

$$e_{\mathrm{CLL}}(\omega,\Omega) = \mathrm{CLL}\left(\hat{H}(\omega,\Omega\right) - \mathrm{CLL}\left(H(\omega,\Omega)\right)$$

is depicted for different methods with $N = 3$ and a cut-on frequency of 2 kHz. In Fig. 3 the CLL error statistics for different Ambisonic orders are depicted for octave bands around 5 and 10 kHz, respectively. These results show that the proposed method consistently outperforms the two methods recently proposed in [2] and [3].

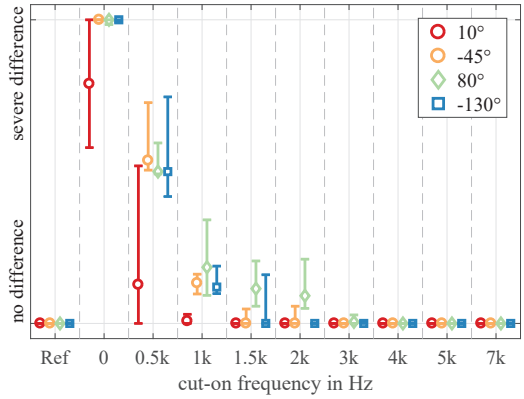


Figure 4: Median an 95% confidence interval of perceived difference ratings for drums as source signal. The cut-on frequency of the modified HRTFs is indicated by the x-ticks.

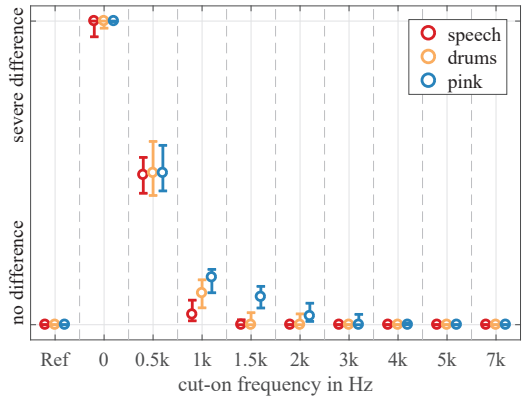


Figure 5: Median an 95% confidence interval of pooled (for all four test directions) ratings per source signal. The cut-on frequency of the modified HRTFs is indicated by the x-ticks.

Optimal Cut-On Frequency

The present method and the method proposed in [3] rely on the assumption, that the relative perceptual importance of IPDs is negligible at high frequencies (compared to interaural level differences and perceived timbre). While this assumption is well motivated, it is not clear how to choose the cut-on frequency ω_c above which IPDs can be disregarded. We therefore conducted a listening experiment using a modified HRTF set as defined in (7). The experiment compared the modified HRTFs with cut-on frequencies $\omega_c = \{0, 0.5, 1, 1.5, 2, 3, 4, 5, 7\}$kHz in a MUSHRA-like procedure against a reference HRTF. Participants were asked to rate the perceived overall difference on a scale from *no audible difference* to *severe difference*. The presented test signals were continuously looped and participants were allowed to seamlessly switch between signals in real-time as often as desired. Overall, three source signals (speech, drum loop, and pulsed pink noise with 150ms hann-windowed ramps) and four source directions $\phi_q = \{10°, -45°, 80°, -130°\}$ were tested in a random sequence. The median and 95% confidence interval of ratings (7 participants, all male, average age 32 years) per source direction for the drum signal are depicted in Fig. 4. While for frontal directions cut-on frequencies above $\omega_c > 1.5$kHz are not significantly different to the reference (Kruskal Wallis test, $p = 0.51$), a minimum cut-on frequency of $\omega_c > 2$kHz is required for lateral directions ($p = 0.84$). Results for speech and pulsed noise showed similar direction-dependent behavior and are therefore not depicted here.

Results of the pooled data per source signal (all four directions) are presented in Fig. 5. The lowest cut-on frequencies which are not significantly different to the reference are $\omega_c = 2$kHz ($p = 0.164$), $\omega_c = 3$kHz ($p = 0.326$), and $\omega_c = 4$kHz ($p = 0.413$) for speech, drums, and pulsed noise, respectively. The increased phase-sensitivity for pulsed noise is explained by onset ITD and envelope ITD evaluation. However, for natural signals (speech, drums) a cut-on frequency as low as $\omega_c = 2$kHz is considered to be sufficient (well inline with duplex theory [7]).

Conclusion

We proposed a method to design binaural rendering filters for Ambisonic signals based on magnitude-only optimization at high frequencies. It has been shown that errors related to the perceived timbre of order-limited Ambisonic signals can be significantly reduced by disregarding interaural phase differences above some cut-on frequency. In a formal listening experiment we found, that this cut-on frequency can be as low as 2 kHz for most signals.

References

[1] B. Bernschütz, A. Vazquez Giner, C. Pörschmann, and J. Arend, "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 972–983, 2014.

[2] 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, 2017.

[3] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural rendering of Ambisonic signals by HRIR time alignment and a diffuseness constraint," *J. Acoust. Soc. Am., submitted*, 2018.

[4] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *J. Acoust. Soc. Am.*, vol. 111, no. 5, p. 2219, 2002.

[5] E. J. Candès, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming," *Comm. on Pure and App. Math.*, vol. 66, no. 8, pp. 1241–1274, 2013.

[6] Z.-Q. Luo, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinit relaxation of quadratic optimization problems," *IEEE Signal Processing Magazine*, no. May, pp. 20–34, 2010.

[7] W. M. Hartmann, B. Rakerd, Z. D. Crawford, and P. X. Zhang, "Transaural experiments and a revised duplex theory for the localization of low-frequency tones," *J. Acoust. Soc. Am.*, vol. 139, no. 2, p. 968, 2016.

5 Binaural Rendering with Measured Room Responses

5.1 BRIR Synthesis using First-Order Microphone Arrays

This work was published as:

M. Zaunschirm, M. Frank, and F. Zotter. (2018). BRIR synthesis using first-order microphone arrays. *Convention of the Audio Eng. Soc. 144*, Milano, pages 1–10.

The idea and concept of this article were outlined by all authors. I, as first author wrote the original draft of the manuscript with periodical contributions from the second and third author. I did the programming and the objective evaluation of the various synthesis methods. The perceptual evaluation was outlined and discussed by all authors, while I did the programming and rendering, the second author and I conducted the experiment, and did the statistical analysis.

Audio Engineering Society

Convention Paper

Presented at the 144th Convention
2018 May 23 – 26, Milan, Italy

BRIR synthesis using first-order microphone arrays

Markus Zaunschirm, Matthias Frank, and Franz Zotter

Institute of Electronic Music and Acoustics, University of Music and performing Arts, Graz

Correspondence should be addressed to Markus Zaunschirm (zaunschirm@iem.at)

ABSTRACT

Both the quality and immersion of binaural auralization benefit from head movements and individual measurements. However, measurements of binaural room impulse responses (BRIRs) for various head rotations are both time consuming and costly. Hence for efficient BRIR synthesis, a separate measurement of the listener-dependent part (head-related impulse responses, HRIR) and the room-dependent part (RIR) is desirable. The room-dependent part can be measured with compact first-order microphone arrays, however the inherent spatial resolution is often not satisfying. Our contribution presents an approach to enhance the spatial resolution using the spatial decomposition method in order to synthesize high-resolution BRIRs that facilitate easy application of arbitrary HRIRs and incorporation of head movements. Finally, the synthesized BRIRs are compared to measured BRIRs.

1 Introduction

The binaural auralization of acoustic environments or the virtualization of an acoustic scene is typically achieved by convolving a source signal with measured binaural room impulse responses (BRIRs) (individual or using an artificial head) and a subsequent playback over headphones [1, 2]. Typically, the BRIR measurements are both time consuming and costly as the artificial head or each future listener has to be carried to the room to be measured. Moreover, BRIRs have to be measured for various head-rotations in order to allow for dynamic binaural rendering.
Dynamic binaural rendering of object-based audio with BRIRs typically requires a switching of the IRs [3], while rendering using a BRIR represented in Ambisonics (scene-based audio) allows for simple rotation by a frequency-independent matrix multiplication [4, 5] while keeping the IRs of the binaural renderer static.
An efficient and versatile method for BRIR synthesis requires a separation in a listener-dependent and a room-dependent part [6, 7]. The listener-dependent part is typically described by high-resolution far-field head-related impulse responses (HRIRs), and the room-dependent part contains the spatio-temporal information at the listening position. The most efficient way (little hardware effort) to capture the room-dependent part employs a first-order microphone array. However, it has been shown in [8] that the first-order representation of RIRs is not sufficient to preserve decorrelation in the reverberation and results in decreased perceived spatial depth for loudspeaker playback.
Higher directional resolution can be achieved by higher-order microphone arrays (e.g. mh acoustics eigenmike) or by directional sharpening of the first-order RIRs using the spatial decomposition method (SDM) [9].

In this method, directional sharpening is achieved by assigning a discrete direction to each sample of the omnidirectional RIR, where the directions are estimated using e.g. the pseudo-intensity vector (PIV) method [10]. SDM allows for a re-encoding of the measured RIR to any desired Ambisonics order. However, directional sharpening of the RIRs leads to an unnatural increase of the reverberation time at high frequencies, especially when using high encoding orders and thus, an order-dependent spectral correction is necessary [8, 11].
In this paper we present an efficient method for synthesizing BRIRs using measured first-order/4-channel (tetrahedral microphone array, see Fig. 1(b)) RIRs followed by directional sharpening and a convolution with pre-measured high-resolution HRIRs of an artificial head [12]. In a listening experiment, the synthesized BRIRs are compared to BRIRs, which were measured with the same artificial head (KU 100, see Fig. 1(a)). Experiments are conducted for different rooms (different reverberation times), various source positions, and evaluate the perceptual attributes of source width, source distance and diffuseness. The tested BRIR synthesis methods include measured first-order RIRs, and directionally sharpened RIRs using different Ambisonics orders, as well as mapping to the nearest direction available in the HRIR grid.

(a) Neumann *KU 100* artificial head.

(b) TSL *ST450* 4-channel Ambisonic B-format microphone array.

Fig. 1: Measurement equipment used.

2 Measurement-Based BRIR Synthesis

There are various approaches for obtaining BRIRs from measured RIRs of a compact, first-order spherical microphone array, see Fig. 2. The two-staged process consists of (i) extraction of directional information and recombination in the *Directional RIR* stage followed by a (ii) combination of the room-dependent and the listener-dependent part (HRIRs) in the *Rendering stage*.

2.1 Directional RIR

For efficient and low-effort measurements we suggest using a compact first-order tetrahedral spherical microphone array, albeit any 3D array configuration can be used. We define the discrete-time single-input-multiple-output (SIMO) RIRs $h(t)$, $x(t)$, $y(t)$, $z(t)$ as the responses of the four-channel output of the *ST450* array (see Fig. 1(b)) after deconvolution by the measurement signal. The four channels (B-format) correspond to a RIR measurement with four independent directivity patterns: omnidirectional for $h(t)$, figure-of-eight in x for $x(t)$, y for $y(t)$, and z-direction for $z(t)$. Similar to the SDM approach [9], we assign a DOA $\boldsymbol{\theta}(t)$ to each discrete-time sample t of the RIR $h(t)$.

For the DOA estimation, we suggest using the pseudo-intensity vector (PIV) approach in the frequency range from 200 Hz and 3 kHz where the directivity patterns of the microphone can be regarded as coincident and clean [10]. We perform a zero-phase band limitation (e.g. by MATLAB's filtfilt) denoted by $F_{200-3\mathrm{k}}$ and a zero-phase smoothing F_L of the resulting PIV using a moving-average time window on the interval $[-L/2;L/2]$ for $L = 16$ around each sample and get the DOA estimate

$$\boldsymbol{\theta}(t) = \frac{\tilde{\boldsymbol{\theta}}(t)}{\|\tilde{\boldsymbol{\theta}}(t)\|}, \quad \text{with} \tag{1}$$

$$\tilde{\boldsymbol{\theta}}(t) = F_L\left\{ F_{200-3k}\{h(t)\}\, F_{200-3k}\left\{ \begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix} \right\} \right\}$$

$\boldsymbol{\theta}(t)$ as Cartesian unit vector.

2.2 Rendering Methods

In order to obtain the synthesized BRIRs, the *directional RIR* is combined with the listener-dependent part, the HRIRs. Now let us consider an arbitrary HRIR set

$$\boldsymbol{A}(t) = [\boldsymbol{a}_1(t), \cdots, \boldsymbol{a}_p(t), \cdots, \boldsymbol{a}_P(t)], \tag{2}$$

$$\boldsymbol{a}_p(t) = [a_p^l(t), a_p^r(t)]^T, \tag{3}$$

where $(\cdot)^{l,r}$ indicates the left and right ear, $(\cdot)^T$ is the transpose operator, the index p indicates the $p-$th direction defined by the normalized Cartesian direction vector $\boldsymbol{\theta}_p = [x_p, y_p, z_p]^T$ of the HRIR sampling grid, and P is the total number of HRIRs.

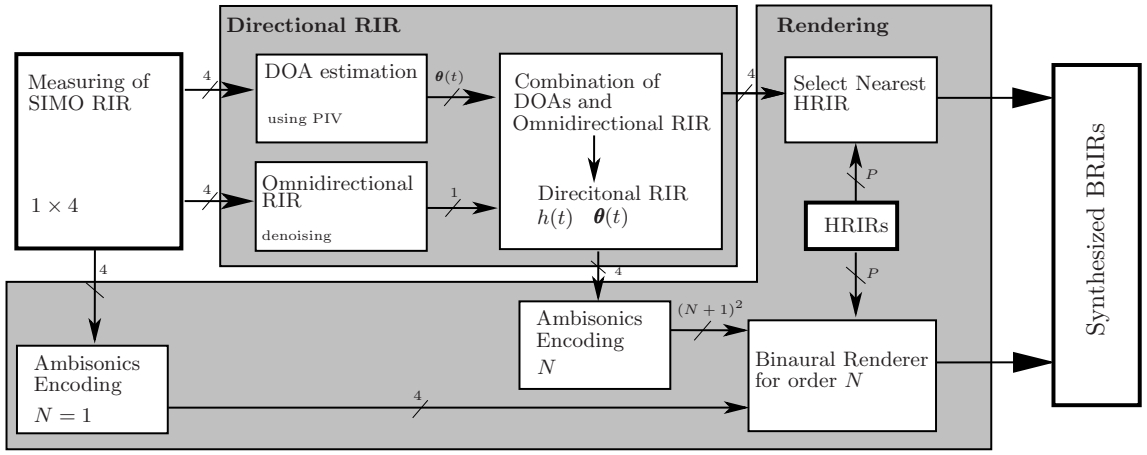


Fig. 2: Block diagram of BRIR synthesis. The measured SIMO RIRs are used for extracting an omnidirectional RIR and corresponding DOA estimates at the receiver location. In the rendering stage the directional RIR is either represented in the Ambisonics domain (see Eq. (8)) and rendered via a state-of-the-art Ambisonics renderer or directly rendered by selecting a HRIR from a pre-measured data set, see Eq. (4).

The synthesized BRIRs are either obtained by (i) time-variant selection of the HRIR direction nearest to the estimated DOA $\boldsymbol{\theta}(t)$, or (ii) binaural rendering of the directional RIR represented in Ambisonics.

2.2.1 Nearest Neighbor Selection

With the unit vector $\boldsymbol{\theta}(t)$, the BRIR synthesis by selection of the nearest-neighbor HRIR pair $a_p^{l,r}(t)$ is

$$BRIR_{NN_P}^{l,r}(t) = \sum_{\tau=0}^{T-1} h(\tau) a_{\tilde{p}(\tau)}^{l,r}(t-\tau), \qquad (4)$$

$$\tilde{p}(t) = \arg\max_p \ \boldsymbol{\theta}_p^T \boldsymbol{\theta}(t), \qquad (5)$$

where $\tilde{p}(t)$ corresponds to the HRIR sampling grid index that is closest to the DOA estimation at discrete time index t, and T is the length of $h(t)$.

2.2.2 Ambisonics

The omnidirectional impulse response $h(t)$ is mixed using the time-dependent direction-of-arrival vector $\boldsymbol{\theta}(t)$ to get a first version of a higher-order Ambisonics room impulse response

$$\tilde{h}_{nm}(t) = Y_n^m[\boldsymbol{\theta}(t)]\, h(t), \qquad (6)$$

where $Y_n^m(\boldsymbol{\theta})$ are the N3D-normalized, real-valued spherical harmonics of order n and degree m evaluated at the direction $\boldsymbol{\theta}$, and N is the maximum order.

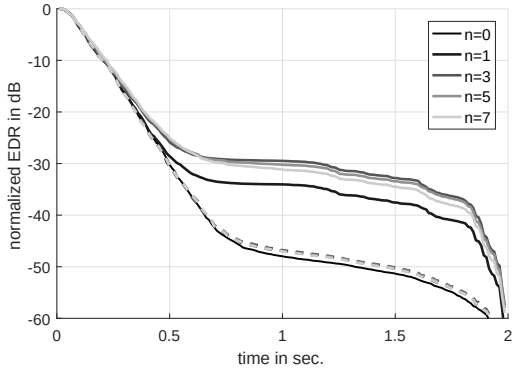


Fig. 3: Energy decay relief (EDR) in a third-octave band with center frequency of 2 kHz. Solid and dashed lines indicate the order partitioned EDR before and after equalization as defined in Eqs. (6) and (7), respectively.

A fast variation of the direction of arrival $\boldsymbol{\theta}(t)$ causes strong amplitude modulation and destroys narrow-band spectral content in $\tilde{h}_{nm}(t)$ by spectral whitening; typically, the longer low-frequency reverberation tails are hereby mixed towards higher frequencies, causing unnaturally long reverberation there [8, 11], cf. solid lines in Fig. 3. Therefore, the response $\tilde{h}_{nm}(t)$ needs spectral correction. To this end, third-octave filtering $\tilde{h}_{nm}(t) = \sum_b F_b\{\tilde{h}_{nm}(t)\}$ is useful, where the b-th subband signal with center frequency f_b is obtained by perfectly reconstructing zero-phase filters F_b.

The time-variant envelope $w_n^b(t)$ accomplishes spectral correction of the sub-band response

$$F_b\{h_{nm}(t)\} = F_b\{\tilde{h}_{nm}(t)\}\, w_n^b(t), \quad (7)$$
$$\text{with } w_n^b(t) = \sqrt{\tfrac{2n+1}{4\pi}} \sqrt{\frac{F_T\left\{F_b\{h(t)\}^2\right\}}{\sum_m F_T\left\{F_b\{h_{nm}(t)\}^2\right\}}},$$

and an averaging time T (e.g. 10 ms), as derived in appendix A. Dashed lines in Fig. 3 show corrected energy decays of higher-order Ambisonic RIRs for the third octave $f_b = 2$ kHz and the orders $n = \{1,3,5,7\}$.

From $h_{nm}(t) = \sum_b F_b\{h_{nm}(t)\}$ the BRIRs for an order N are synthesized by

$$BRIR_{A_N}^{l,r}(t) = \sum_{\tau=0}^{T-1} \sum_{n=0}^{N} \sum_{m=-n}^{n} b_{nm}^{l,r}(\tau) h_{nm}(t-\tau), \quad (8)$$

where $b_{nm}^{l,r}(t)$ is any state-of-the-art FIR binaural Ambisonic renderer (e.g. [13]) of the length T, or the one favored here that was defined in [14]. Its frequency-dependent time-alignment of the HRIR set and a diffuse-field constraint can significantly improve both the coloration as well as localization accuracy of binaurally rendered Ambisonic signals represented by practical orders $N < 7$.

3 Evaluation

The proposed BRIR synthesis methods are compared and evaluated via both technical measures including the reverberation time (T30), early decay time (EDR), clarity index (C80), apparent source width (ASW) defined as $1 - IACC_E$, and a listening experiment against a reference BRIR.

The reference BRIRs are recorded in real rooms between a single *Genelec 8020* loudspeaker and the *KU 100* artificial head using the exponentially swept sine method [15]. All SIMO RIRs of the *ST450*, which are the basis of the synthesized BRIRs (see Fig. 2), are measured with the same source, at the same position, and using the same excitation signal.

Overall the test conditions include (i) three different rooms at the IEM Graz, and (ii) two different directions or source and receiver distances at a fixed source- and ear-height of 1.3m. In all measurements the source is directed towards the artificial head/microphone array.

The measured rooms, directions ϕ (azimuth angle with origin in center of the artificial head, and positive x-axis through the nose) and source-receiver distance r are

- Production studio (PS): volume 127 m^3, base area 42 m^2, $T_{60} \approx 0.4$ s. Directions: $\phi = 0°$, and $\phi = 90°$. Source distance $r = 2.3$ m.
- CUBE (CU): volume 620 m^3, base area 130 m^2, $T_{60} \approx 0.7$ s. Directions: $\phi = 0°$, and $\phi = 90°$. Source distances $r = 2.3$ m, and $r = 4$ m.
- Corridor (CO): volume 210 m^3, base area 64 m^2, $T_{60} \approx 1.4$ s. Direction $\phi = 0°$. Source distance $r = 10$ m.

For synthesis we used the omni-directional (W-channel) of the *ST450* output and for rendering a far-field HRIR data set of the *KU 100* measured at overall $P = 2702$ sampling points [12]. The tested synthesis methods include

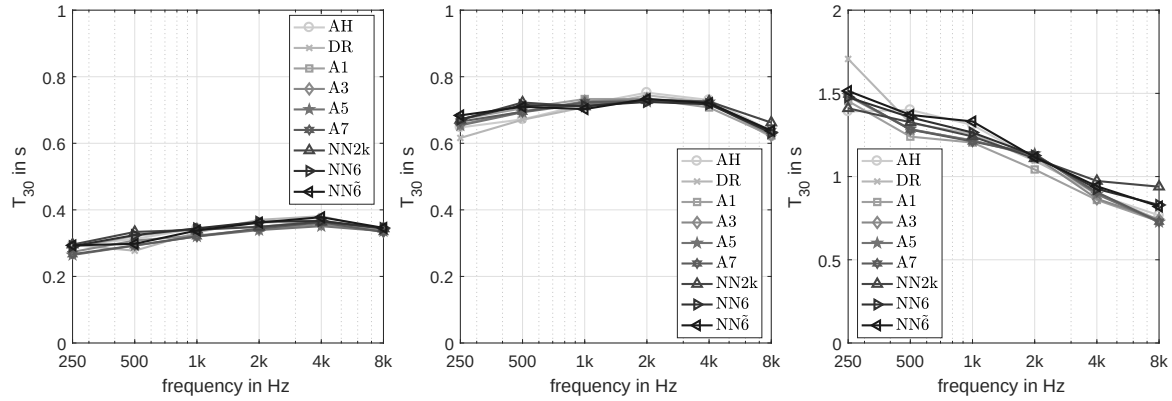
- Direct rendering (DR) of the measured B-format RIRs, see Eq. (8).
- Nearest neighbor rendering with the entire HRIR set (NN2k), with a subset consisting of 6 HRIRs at the front, left, back, right, top, and bottom (NN6), and a subset of 6 HRIRs at front-left, back-left, back-right, front-right, top, and bottom ($\widetilde{\text{NN6}}$), see Eq. (4).
- Rendering using the directionally sharpened RIR with orders $N = \{1,3,5,7\}$. The corresponding synthesized BRIRs are abbreviated as A1, A3, A5, A7, respectively.

3.1 Technical Measures

In the following section, the measured reference and synthesized BRIRs are analyzed in terms of technical measures as defined in [16]. For all measures which require a single channel input, the BRIRs of the left and right ear are averaged and parameters are calculated with the Lundeby method [17] which is employed in the AcMus toolbox [18]. As the RIRs are measured for multiple directions and source distances, the parameters are averaged to give a single value per room.

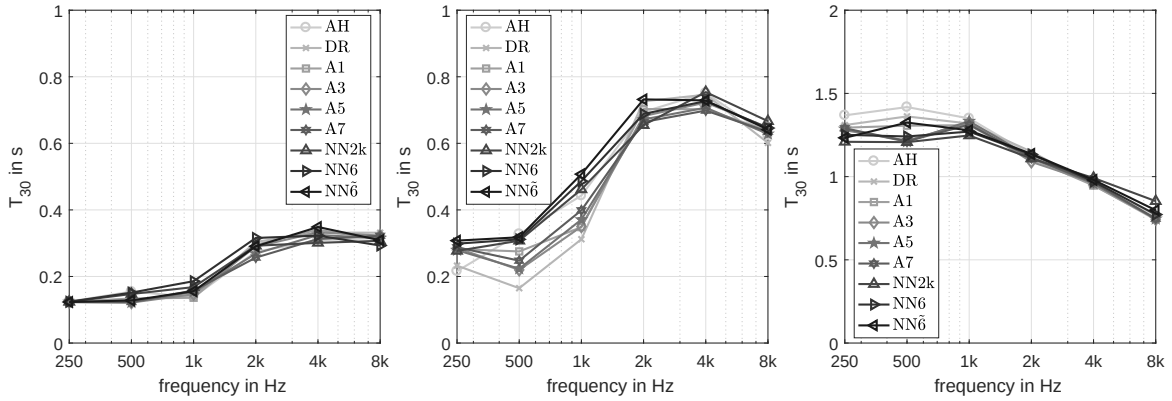
3.1.1 Reverberation Time

The typical measure for the energy decay rate in a room is the reverberation time, which is typically calculated via the Schroeder backwards integration in octave bands between 250 Hz and 8 kHz [16]. The resulting T_{30} values for all synthesis methods and for each of the measured rooms are depicted in Fig. 4. As expected, little variation is observed across the rendering methods as the processing not alters the energy decay of the measured omnidirectional response.



(a) Production Studio. (b) CUBE. (c) Corridor.

Fig. 4: Reverberation time T_{30} in octave bands between 250 Hz and 8 kHz for the three evaluated rooms.



(a) Production Studio. (b) CUBE. (c) Corridor.

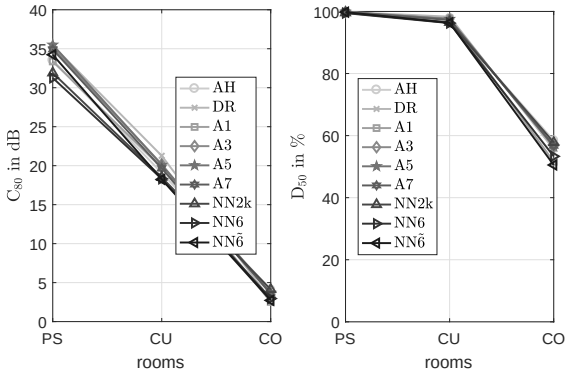
Fig. 5: Early decay time (EDT) in octave bands between 250 Hz and 8 kHz for the three evaluated rooms.

3.1.2 Early Decay Time

According to [16] the early decay time (EDT) is considered to be a more appropriate technical measure for the reverberance of a room than the reverberation time. As the EDT is based on the slope of the 10dB drop in the normalized energy decay curve (EDC), early reflections contribute more significantly to the EDT when compared to the reverberation time. While the EDTs for PS and CO are well aligned with the reference (AH), a deviation can be observed for CU at the lower octave bands, see Fig. 5. In [16] the JND for EDT is quantified as 5%, however it is pointed out in [19] that the JND is highly dependent on the source signal and that JNDs between 25% can be expected.

3.1.3 Clarity and Definition

The technical measures of speech intelligibility and transparency of music as defined in [16] are the definition (D_{50}) and clarity (C_{80}), respectively. Results for the reference and synthesized BRIRs are shown in Fig. 6. For most conditions and rooms the results show little deviation to the reference (2dB clarity change in CU, and 5% definition change in CO). As expected, clarity and definition are higher in rooms with lower T_{30}, cf.. Fig. 4. According to the JNDs given in [16] (1dB for clarity, and 5% for definition) the worst case deviations lie between 1-3 JNDs (NN6̃ in CU).



(a) Clarity index. (b) Definition.

Fig. 6: Single number clarity C_{80} and definition D_{50} for all rooms (averaged between 500 Hz and 1 kHz octave band).

3.1.4 Apparent Source Width

In [20] the apparent source width ASW is defined as *apparent auditory width of the sound field created by a performing entity as perceived by a listener* and modeled via a measure related to the interaural cross-correlation coefficient (IACC). Here the IACC is calculated in the integration interval between [0,80]ms and $(1-IACC_{[0,80]ms})$ is depicted in Fig. 7. It can be seen that DR, NN6, and NN6̃ show significant deviations from the reference (JND for IACC is defined frequency-independent as 0.075 in [16], although it has been shown in [21] that the JNDs strongly depend on the reference condition and range from $0.08-0.35$).

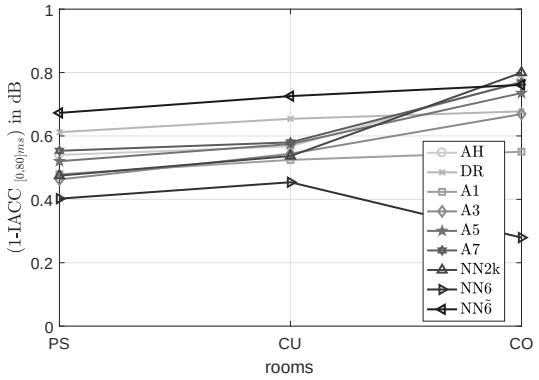


Fig. 7: $(1-IACC_{[0,80]ms})$ for all rooms and BRIR synthesis methods.

3.2 Listening Experiment

For the common technical measures, which are based on energy decay or ratios, no clear difference is observed between the reference (AH) and synthesized BRIRs, see Sec. 3.1. In order to evaluate audible differences, a listening experiment was conducted. The experiment compared the above-mentioned synthesis/rendering methods (DR, A1, A3, A5, A7, NN2k, NN6, and NN6̃) in a MUSHRA-like [22] procedure against the artificial head AH as a reference.
Note that global timbral deviations between the reference and the synthesized BRIRs, which occur due to spatial aliasing [23], array imperfections (encoding) and microphone frequency responses, are equalized with a single global minimum-phase equalization filter for all synthesis methods and both ears.
The comparison evaluated 3 attributes that seemed reasonable from informal listening by the authors:

- Width: how wide is the source spread and how blurry is the localization of the direct sound?
- Diffuseness: how evenly is the reverberation distributed, are distinct spatial areas audible?
- Distance: how far is the source perceived?

As the localization of the direct sound was only altered by the NN6̃ renderer, this attribute was omitted.
The experiment included all 7 room conditions (PS ($r = 2.3$ m) and CU ($r = \{2.3, 4.0\}$ m) for source directions of $\phi = 0°, 90°$ and CO ($r = 10$ m) for $\phi = 0°$). It was divided into 3 parts, one for each attribute. The parts were performed in random order and within each part, room conditions and rendering methods were also randomized. The source signal were the first 5 seconds of the EBU female German speech recording [24]. Playback employed equalized AKG K702 headphones powered by an RME Multiface.
6 listeners with experience in spatial audio (all male, average age 33 years) participated in the experiment and it took them on average 56 minutes.

3.2.1 Results

For each attribute, room condition, and listener, the answers were normalized to a maximum absolute difference of 1 to the reference. The results for width and diffuseness could be summarized over all 7 room conditions. However for distance, their was a clear separation into two groups: one with frontal direct sound (4 conditions) and one with direction sound from the side (3 conditions).

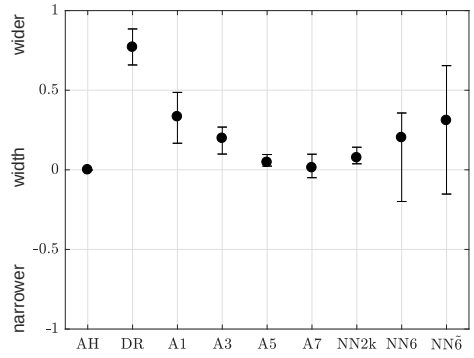


Fig. 8: Median values and corresponding 95% confidence intervals for perceived width, summarizing all 6 listeners and 7 room conditions.

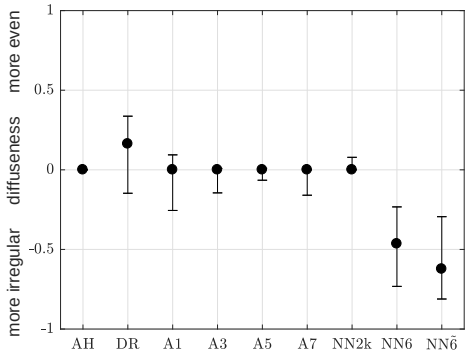


Fig. 10: Median values and corresponding 95% confidence intervals for perceived diffusity, summarizing all 6 listeners and 7 room conditions.

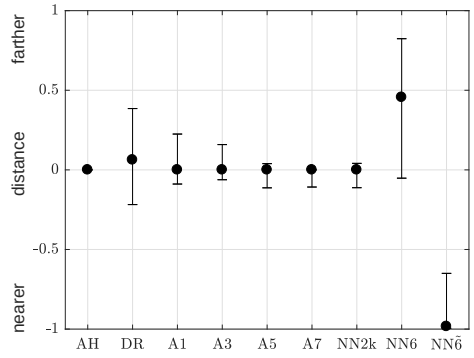


Fig. 9: Median values and corresponding 95% confidence intervals for perceived distance, summarizing all 6 listeners and 4 room conditions with frontal direct sound (PS ($r = 2.3\,\mathrm{m}$), CU ($r = \{2.3, 4.0\}\,\mathrm{m}$), and CO ($r = 10.0\,\mathrm{m}$) for $\phi = 0°$).

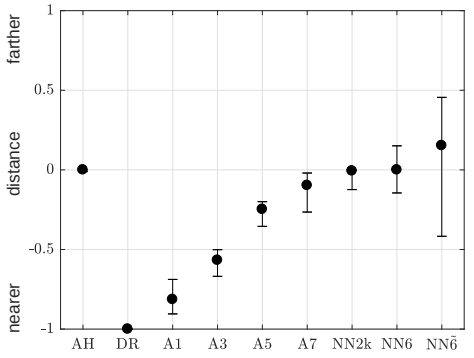


Fig. 11: Median values and corresponding 95% confidence intervals for perceived distance, summarizing all 6 listeners and 3 room conditions with lateral direct sound (PS ($r = 2.3\,\mathrm{m}$), CU ($r = \{2.3, 4.0\}\,\mathrm{m}$) for $\phi = 90°$).

Direct rendering of the 1st-order RIR (DR) was perceived significantly wider ($p \leq 0.018$) than all other rendering methods and the reference AH, cf. Fig. 8. Width decreases with the Ambisonics order, so that A5 and A7 ($p > 0.11$) are not distinguishable from AH. Although from the NN renderers only NN2k is significantly wider than the reference AH ($p = 0.026$), NN6 and NN6̃ exhibit an undesirable large spread.

NN6 and NN6̃ yield significantly less even diffuseness than all other renderers and the reference AH ($p \leq 0.015$), whereas both are not significantly different, see Fig. 10. All remaining renderers are not distinguishable from AH ($p > 0.12$) and among themselves ($p > 0.065$).

As Fig. 9 shows, NN6̃ yields the smallest distance of all renderers ($p << 0.001$) for frontal direct sound. N6 yields farther results ($p \leq 0.016$) than most renderers, except for DR and A1 ($p > 0.067$). DR, all Ambisonic renderers, and NN2k are indistinguishable from the reference AH ($p > 0.15$).

For direct sound from the side, cf. Fig. 11, the perceived distance significantly increases from DR to A1 to A3 to A5 ($p \leq 0.01$). Further increasing of the order to 7 does not significantly increase the distance ($p = 0.071$), however A7 is the only Ambisonic renderer that is indistinguishable from the reference AH ($p = 0.093$). Moreover, all NN renderers are not distinguishable from the reference AH ($p > 0.33$), however NN6̃ again exhibits an undesirable large spread.

3.3 Discussion

While from the technical measures only the IACC indicates differences between the tested renderers, the listening experiment revealed significant differences for all evaluated attributes. Except for diffuseness and distance of frontal sources, the direct binaural rendering of the 1st-order RIR (DR) significantly deviates from the measured BRIR (AH). Unsurprisingly, the deviation decreases with the Ambisonic order of the directionally sharpened RIRs. With an order of 7, the synthesized BRIR is indistinguishable from the measured BRIR. Similarly good results are obtained for NN2k. In most cases, the BRIRs synthesized by the nearest neighbor renderers with only 6 directions (NN6 and NN$\tilde{6}$) largely differ from the measured BRIR and their results have a large spread. Their results also strongly depend on whether the direction of the direct sound coincides with the directions of the selectable HRIRs. Moreover, the sparse mapping to 6 directions also impairs the evenness of the reverberation, resulting in reduced diffuseness.

4 Conclusion

In this contribution we presented an efficient two-staged measurement-based BRIR synthesis method, which allows for subsequent incorporation of arbitrary HRIRs. In the first stage, the measured SIMO RIRs of a compact tetrahedral microphone array are used for both the extraction of the omnidirectional RIR and time-variant DOA estimation, similar to SDM [9]. In a rendering stage the omnidirectional RIR and DOA estimates are either used for (i) nearest neighbor rendering to directions available in the HRIR set or (ii) Ambisonic rendering of a directionally sharpened RIR. Listening experiments compared the synthesized BRIRs against the measured reference BRIRs for 3 rooms with different acoustic characteristics and source positions. Using Ambisonic rendering, an order of 7 yields results that are indistinguishable from the reference in terms of distance, width, and diffuseness. Reduction of the Ambisonic order increases the deviation from the reference. Interestingly, the sharpened 1st-order Ambisonic rendering outperforms the direct rendering of the measured 1st-order RIRs.
The nearest neighbor rendering with $P = 2702$ HRIR directions yields similar results to 7th-order Ambisonics; results for using $P = 6$ directions show strong deviations from the reference and undesirably large spread and are therefore not recommended.

Moreover, the higher-order representation of the directional RIR allows for rotation of the acoustic scene via a frequency-independent matrix multiplication to account for head movements prior to rendering with a static set of filters for BRIR synthesis.

References

[1] Wightman, F. L. and Kistler, D. J., "Headphone simulation of free field listening I: stimulus synthesis," *J. Acoust. Soc. Am.*, 85(1989), pp. 858–867, 1989.

[2] Møller, H., "Fundamentals of binaural technology," *Applied Acoustics*, 36(3-4), pp. 171–218, 1992, ISSN 0003682X, doi:10.1016/0003-682X(92)90046-U.

[3] Engdegard, J., Resch, B., Falch, C., Hellmuth, O., Hilpert, J., Hoelzer, A., Breebaart, J., Koppens, J., Schuijers, E., and Oomen, W., "Spatial Audio Object Coding (SAOC) – The Upcoming MPEG Standard on Parametric Object Based Audio Coding," *124th AES Convention*, 2008.

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

[5] Pinchon, D. and Hoggan, P. E., "Rotation matrices for real spherical harmonics: General rotations of atomic orbitals in space-fixed axes," *Journal of Physics A: Mathematical and Theoretical*, 40(7), pp. 1597–1610, 2007, ISSN 17518113, doi:10.1088/1751-8113/40/7/011.

[6] Pörschmann, C. and Wiefling, S., "Perceptual Aspects of Dynamic Binaural Synthesis based on Measured Omnidirectional Room Impulse Responses," *International Conference on Spatial Audio*, (December 2016), 2015.

[7] Menzer, F., *Binaural Audio Signal Processing Using Interaural Coherence Matching*, Ph.D. thesis, 2010.

[8] Frank, M. and Zotter, F., "Spatial impression and directional resolution in the reproduction of reverberation," in *Proc. DAGA*, pp. 1304–1307, Aachen, 2016.

[9] Tervo, S., Pätynen, J., Kuusinen, A., and Lokki, T., "Spatial decomposition method for room impulse responses," *Journal of the Audio Engineering Society*, 61(1/2), pp. 17–28, 2013.

[10] Jarrett, D. P., Habets, E. A. P., and Naylor, P. A., "3D Source localization in the spherical harmonic domain using a pseudointensity vector," *European Signal Processing Conference*, (April), pp. 442–446, 2010, ISSN 22195491.

[11] Zaunschirm, M., Baumgartner, C., Schörkhuber, C., Frank, M., and Zotter, F., "An Efficient Source-and-Receiver-Directional RIR Measurement Method," in *Fortschritte der Akustik AIA-DAGA 2017*, pp. 1343–1346, Kiel, 2017.

[12] Bernschütz, B., "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," *Fortschritte der Akustik – AIA-DAGA 2013*, pp. 592–595, 2013.

[13] Bernschuetz, B., Vazquez Giner, A., Poerschmann, C., and Arend, J., "Binaural reproduction of plane waves with reduced modal order," *Acta Acustica united with Acustica*, 100(5), pp. 972–983, 2014, ISSN 16101928, doi:10.3813/AAA.918777.

[14] Zaunschirm, M., Schoerkhuber, C., and Hoeldrich, R., "Binaural rendering of Ambisonic signals by HRIR time alignment and a diffuseness constraint," *J. Acoust. Soc. Am., submitted*, 2018.

[15] Farina, A., "Simultaneous measurement of impulse response and distortion with a swept-sine technique," *Proc. AES 108th conv, Paris, France*, (I), pp. 1–15, 2000, doi:10.1109/ASPAA.1999.810884.

[16] BS EN ISO 3382-1:2009, "Acoustics - Measurement of room acoustic parameters. Part 1:Performance spaces," *British Standard*, pp. 1 – 26, 2009, ISSN 1098-6596, doi:10.1017/CBO9781107415324.004.

[17] Lundeby, A., Vigran, T. E., Bietz, H., and Vorländer, M., "Uncertainties of measurements in room acoustics," *Acta Acustica united with Acustica*, 81(4), pp. 344–355, 1995.

[18] Queiroz, M., Iazzetta, F., Kon, F., Gomes, M. H. a., Figueiredo, F. L., Masiero, B., Ueda, L. K., Dias, L., Torres, M. H. C., and Thomaz, L. F., "AcMus: an open, integrated platform for room acoustics research," *Journal of the Brazilian Computer Society*, 14(3), pp. 87–103, 2008, ISSN 0104-6500, doi:10.1007/BF03192566.

[19] Meng, Z., Zhao, F., and He, M., "The just noticeable difference of noise length and reverberation perception," *2006 International Symposium on Communications and Information Technologies, ISCIT*, pp. 418–421, 2006, doi:10.1109/ISCIT.2006.339980.

[20] Hidaka, T., "Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls," *The Journal of the Acoustical Society of America*, 98(2), pp. 988–1007, 1995, ISSN 00014966, doi:10.1121/1.414451.

[21] Kim, C., Mason, R., and Brookes, T., "Initial investigation of signal capture techniques for objective measurement of spatial impression considering head movement," in *124th AES Convention, Amsterdam, Netherlands*, volume 7331, 2008, ISBN 9781605602950.

[22] International Telecommunication Union, "ITU-R BS.1534-3, Method for the subjective assessment of intermediate quality level of audio systems," *ITU-R Recommendation*, 1534-3, pp. 1534–3, 2015.

[23] Rafaely, B., Weiss, B., and Bachmat, E., "Spatial aliasing in spherical microphone arrays," *IEEE Transactions on Signal Processing*, 55(3), pp. 1003–1010, 2007, ISSN 1053587X, doi:10.1109/TSP.2006.888896.

[24] European Broadcasting Union, "EBU Tech 3253 - Sound Quality Assessment Material recordings for subjective tests," (September), 2008.

A Spectral Correction of Directionally Sharpend RIRs in Ambisonics

The squared impulse response after SDM upmixing is

$$\tilde{h}^2_{nm}(t) = |Y^m_n[\boldsymbol{\theta}(t)]|^2\, h^2(t), \tag{9}$$

and due to the pseudo-allpass property $\sum_{m=-n}^{n}|Y^m_n(\boldsymbol{\theta})|^2 = 2n+1$ of the orthonormal spherical harmonics, $\int_{\mathbb{S}^2}|Y^m_n(\boldsymbol{\theta})|^2\,\mathrm{d}\boldsymbol{\theta} = 1$, we obtain a relation between energies in the DOA-modulated n^{th}-order and the 0^{th} order impulse response $\tilde{h}^2_{00}(t) = \frac{1}{4\pi}h^2(t)$,

$$\sum_{m=-n}^{n}\tilde{h}^2_{nm}(t) = \tfrac{2n+1}{4\pi}\, h^2(t). \tag{10}$$

For a spectral correction, we observe that this property remains unaffected when summing over T discrete-time instances $t = -\frac{\mathrm{T}}{2},\dots\frac{\mathrm{T}}{2}-1$ around the time instant τ

$$\sum_{t=-\frac{\mathrm{T}}{2}}^{\frac{\mathrm{T}}{2}-1}\sum_{m=-n}^{n}\tilde{h}^2_{nm}(t+\tau) = \frac{2n+1}{4\pi}\sum_{t=-\frac{\mathrm{T}}{2}}^{\frac{\mathrm{T}}{2}-1}h^2(t+\tau), \tag{11}$$

hence Parseval's theorem allows to replace summation over squared discrete-time samples by summation over magnitude squared discrete-frequency Fourier coefficients $\sum_{t=-\frac{\mathrm{T}}{2}}^{\frac{\mathrm{T}}{2}-1}x^2(t) = \sum_{k=0}^{\mathrm{T}-1}|X(k)|^2$. Consequently, the above relation for the energy within the order n of the room response also holds in the frequency domain

$$\sum_{m=-n}^{n}\sum_{k=0}^{\mathrm{T}-1}\tilde{|}\tilde{H}_{nm,\tau}(k)|^2 = \frac{2n+1}{4\pi}\sum_{k=0}^{\mathrm{T}-1}|H_\tau(k)|^2, \tag{12}$$

which finally permits to undertake spectral corrections. To correct the spectrally whitened response $\tilde{H}_{nm,\tau}(k)$, we introduce an equalizer $W_{n,\tau}(k)$ and define the spectrally corrected response for a time-offset τ as

$$H_{nm,\tau}(k) = W_{n,\tau}(k)\,\tilde{H}_{nm,\tau}(k). \tag{13}$$

While the equalizer must retain the above equation for the summed energies over k and m, the equalizer should restore the spectral decay of the response $H_\tau(k)$ for all time shifts τ at all discrete frequencies k

$$\sum_{m=-n}^{n}|W_{n,\tau}(k)\,\tilde{H}_{nm,\tau}(k)|^2 = \tfrac{2n+1}{4\pi}|H_\tau(k)|^2, \tag{14}$$

$$|W_{n,\tau}(k)|^2\sum_{m=-n}^{n}|\tilde{H}_{nm,\tau}(k)|^2 = \tfrac{2n+1}{4\pi}|H_\tau(k)|^2.$$

Thus the spectral decay correction

$$|W_{n,\tau}(k)|^2 = \frac{2n+1}{4\pi}\frac{|H_\tau(k)|^2}{\sum_{m=-n}^{n}|\tilde{H}_{nm,\tau}(k)|^2}. \tag{15}$$

can be applied to the room impulse response as a time-variant filter

$$h_{nm}(t+\tau) = \sum_{k=0}^{\mathrm{T}-1}W_{n,\tau}(k)\,H_{nm,\tau}(k)\,e^{\mathrm{i}\frac{2\pi}{\mathrm{T}}kt}. \tag{16}$$

For smooth results, a third-octave analysis is advised and smoothing of the envelope $W_{n,\tau}(k)$ to a temporal envelope $w^b_n(t)$ that can be applied to the third-octave decomposed impulse response, as above. The envelope $w^b_n(t)$ is obtained from the square-root of the gathered energies $\sqrt{\sum_{k:f_k\in[2^{-1/6};2^{1/6}]f_b}W^2_\tau(k)}$ of the bins around the third-octave center frequencies f_b. These are interpolated over all instants t between the analysis time shifts τ (hop size). In the current implementation, gathering of the energies for the band b employs a $\sin^2$ window from f_{b-1} to f_{b+1} to get smooth transitions between the bands.

5.2 Perceptual Evaluation of Variable-Orientation Binaural Room Impulse Response Rendering

This work was published as:

M. Zaunschirm, M. Frank, and F. Zotter. (2018). Perceptual Evaluation of Variable-Orientation Binaural Room Impulse Response Rendering. *Conference of the Audio Eng. Soc.: 2019 AES International Conference on Immersive and Interactive Audio*, York, UK.

The idea and concept of this article were outlined by all authors. I, as first author wrote the original draft of the manuscript with help from the second author and periodical contributions from the third author. I did most of the programming and graphical work, and prepared the samples for the listening experiment. The listening experiment was designed and programmed equally by all authors.The data was analyzed by the second author and me.

Audio Engineering Society

Conference Paper

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

Perceptual Evaluation of Variable-Orientation Binaural Room Impulse Response Rendering

Markus Zaunschirm[1], Matthias Frank[1], and Franz Zotter[1]

[1] *Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz, 8010, Austria*

Correspondence should be addressed to Markus Zaunschirm (zaunschirm@iem.at)

ABSTRACT

In the current effort to improve sound for virtual auditory environments with the upcoming affordable head mounted display solutions, realism and audio quality in head-tracked binaural rendering is again becoming important. While rendering based on static artificial-head measurements achieve high audio quality and externalization, the realism lacks interactivity with changes of the head orientation. Motion-tracked binaural (MTB) has been presented as a head-tracked rendering algorithm for recordings made with circular arrays on rigid spheres. In this contribution, we investigate the algorithm proposed for MTB rendering and adopt it for variable-orientation rendering using binaural room impulse responses (BRIR) measured for multiple, discrete orientations of an artificial head. The experiment in particular investigates the perceptual implications of the angular resolution of the multi-orientation BRIR sets and the time/frequency-resolution of the algorithm.

1 Introduction

Typically, binaural rendering involves a convolution of source signals with measured or modeled head-related impulse responses (HRIRs) or binaural room impulse responses (BRIRs) and playback via headphones [1]. Both, HRIRs and BRIRs implicitly contain the cues that are evaluated by the human auditory system to perceive the direction, distance, source width, envelopment, and spaciousness of a sound, cf. [2, 3].
Often, it is helpful to involve a natural acoustic room and to render BRIRs, as this reduces poor externalization of both individual and non-individual HRIRs, which is also strongly affected by the room divergence effect [4, 5, 6], i.e. violations of acoustical expectations arising from the environment in which one listens to headphones.

If room divergence is not a problem, room impulses measured with a loudspeaker and an artificial head can achieve static rendering of high audio quality and of convincing realism, when used to convolve monophonic or multi-channel audio material to obtain a suitable pair of headphone playback signals [1, 7]. Such a BRIR-based virtualization of loudspeakers in rooms is versatile and is not limited to virtualizing multi-channel loudspeaker setups from studios and concert halls [8, 9, 10], but is also used to document and preserve acousmatic or electroacoustic music sceneries, or sound installations.

Dynamic BRIR rendering: When playing back such virtualizations, dynamic binaural rendering can help to reduce localization ambiguities and poor externalization by involving the natural interaction of the ear

signals with the head rotation of a listener [11, 12, 13]. While rendering of object-based audio typically involves fading or switching of filters (HRIRs or BRIRs) [14], rendering of Ambisonics (scene-based audio) requires no filter switching as the entire sound scene can be rotated by a simple frequency-independent matrix multiplication [15, 16].
Obviously, the processing and head-tracking must be done in real-time to ensure naturalness of the listening experience. The maximal latency of dynamic binaural rendering in order to avoid spatial 'slewing' (source stability) is investigated in [17, 12, 18], and is found to be between $40-60$ms for simple and complex acoustic scenes, respectively.
To accomplish dynamic and interactive head-tracked BRIR rendering, an additional effort (i.e. scalable with the desired resolution) has to be taken on the measurement side: Instead of measuring only one pair of BRIRs (for the left and the right ear) from each loudspeaker for one artificial head orientation, a pair of BRIRs is measured for a set of head orientations using a calibrated turntable, e.g. [10].

State of the art: When the set of available BRIR orientations is coarse, e.g. with orientations separated by more than 3°, a straight-forward filter-switching approach is not sufficient in order to achieve continuous (no spatial artifacts) and robust dynamic BRIR rendering, see [19]. Thus, suitable interpolation strategies as suggested in [20] are required. The suggested strategies include (i) full-bandwidth linear interpolation (e.g. VBAP interpolation), and (ii) dual-band interpolation where the high-frequency (HF) content is either obtained from a fixed microphone, the nearest microphone, or from a spectral interpolation method.
Spectral interpolation blends the magnitude of the HF signals linearly while a suitable phase can be found by spectrogram inversion, e.g. [21]. Variants of the spectral interpolation method and especially the reconstruction of the HF time signal are discussed in [22] and perceptual evaluations are presented in [23], where it is found that using the phase of the nearest BRIR for re-synthesis performs best in terms of audio quality and computational efficiency. However, perceptually optimized parameter settings of the proposed interpolation methods are missing so far.

In this contribution we discuss a dual-band method in the context of real-time BRIR or ear-signal interpolation with minimal latency, and present perceptually optimized parameter settings in Sections 2, and 3, respectively. In Sec. 4 we show the results from listening experiments where listeners were asked to rate the perceptual attributes (i) timbre, (ii) spatialization, and (iii) continuity of the interpolation compared to directly rendered BRIRs measured for 1° angular resolution. Thereby, we compare different interpolation methods and identify grid resolutions which are needed to obtain perceptually appealing binaural experiences.

2 Method

While the proposed method could be used for interpolation of BRIR orientations without any source signal, we discuss interpolation on the basis of ear signals rendered for differently-oriented BRIRs.
Assuming a BRIR set measured for M equi-angular orientations on the horizon (around the Cartesian z-axis), the interpolation method uses the two BRIRs oriented closest to the head orientation of the listener $\varphi(t)$ at the discrete-time index t. With $\Delta\varphi$ as azimuthal resolution, the indices of the two nearest BRIR orientations are

$$m = \left\lfloor \frac{\varphi(t)}{\Delta\varphi} \right\rfloor, \quad m+1 = \left\lceil \frac{\varphi(t)}{\Delta\varphi} \right\rceil, \tag{1}$$

where $\lfloor\cdot\rfloor$, and $\lceil\cdot\rceil$ are the floor and ceil functions, respectively. Per ear, the corresponding two signals $x_m(t)$, $x_{m+1}(t)$ are rendered by convolution using the closest-orientation BRIRs, see Fig. 1.
In a full-bandwidth linear interpolation the resulting ear signal is obtained as

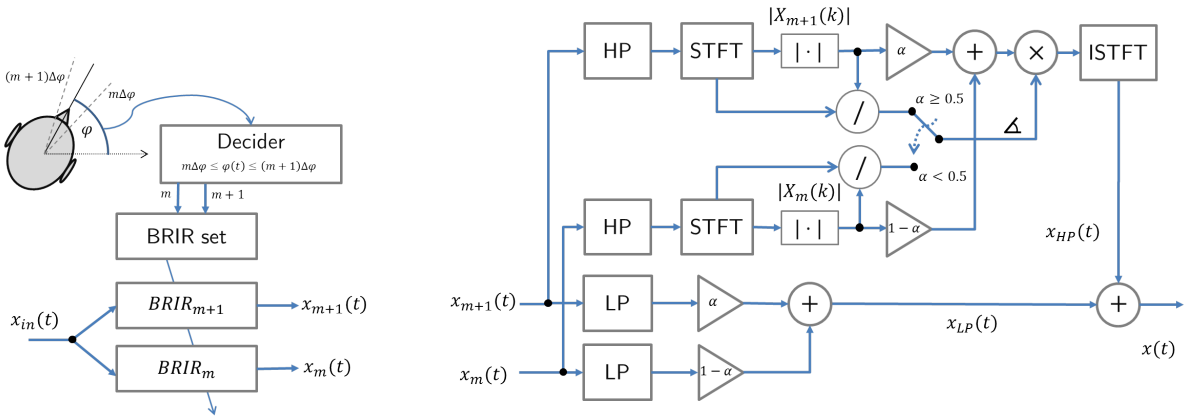
$$x(t) = (1-\alpha)x_m(t) + \alpha x_{m+1}(t), \tag{2}$$

where the linear interpolation weight is obtained by

$$\alpha = \left\lceil \frac{\varphi(t)}{\Delta\varphi} \right\rceil - \frac{\varphi(t)}{\Delta\varphi}. \tag{3}$$

By nature, the linear combination of delayed signals produces comb-filtering which can lead to significant colorations in the resulting signal and thus, we suggest using a dual-band approach as described in [20, 22]. The block diagram of the proposed dual-band interpolation method is depicted in Fig. 1. While the signal in the lower band is processed in the time domain by applying the linear weights, the signal in the high frequency band is obtained by magnitude interpolation

$$X(k) = \{(1-\alpha)|X_m(k)| + \alpha|X_{m+1}(k)|\}\ \mathrm{e}^{i\angle(k)}, \tag{4}$$



(a) Filter-switching and convolution to obtain the two neighbouring ear signals which are closest to the head rotation.

(b) Block diagram of the proposed dual-band ear signal interpolation scheme. The lower band is interpolated linearly in the time domain, while the high-frequency band is interpolated in the short-time Fourier domain. The interpolated high-frequency signal is obtained by a linear magnitude interpolation and the phase for reconstruction is chosen dependent of α (interpolation weight).

Fig. 1: Block diagram for BRIR selection, convolution, and continuous interpolation for one ear in a dynamic binaural rendering scenario. The interpolation scheme can also operate on recorded ear signal.

with the phase of the nearest signal

$$\measuredangle(k) = \begin{cases} \measuredangle_{m+1}(k) & \text{for } \alpha \ge 0.5 \\ \measuredangle_m(k) & \text{else.} \end{cases} \tag{5}$$

Here k is the frequency index of a short-time Fourier transform (STFT) frame, $i^2 = -1$, and $\measuredangle(k)$ is the phase argument. Note that a more elaborate time signal estimation from magnitude-interpolated STFTs uses real-time spectrogram inversion [24], but simplicity is preferred, here. Note that the LP signal $x_{LP}(t)$ has to be delayed in dependence of the STFT settings and the underlying audio block size in order to avoid artifacts in the interpolated signals.

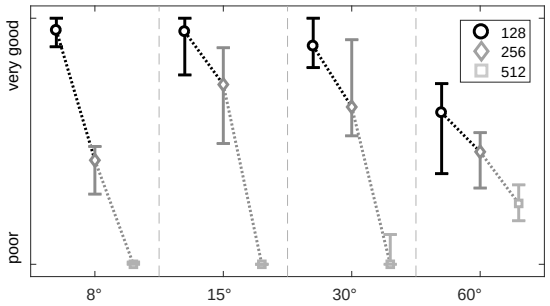
3 Implementation and Settings

With unsuitable grid resolution $\Delta\varphi$, update rate of the *Decider*, or size of the block processing, rendering according to Fig. 1 will be deteriorated by spatial, temporal, or timbral artifacts. While spatial and timbral artifacts are mostly determined by the grid resolution, interpolation method and cross-over frequency, temporal artifacts are mostly due to the block size, e.g. [22]. In our implementation (in *pure data* [1]) the frequency crossover at f_c is composed of 4-th order Linkwitz-Riley filters, as they provide all-pass characteristics and a smoothly changing phase response. In the short-time block processing with block size N and hop size $L = N/2$, a half-wave window (as square root of a Hann window) is applied at both the analysis and synthesis stage to reduce cyclic artifacts at the block boundaries.

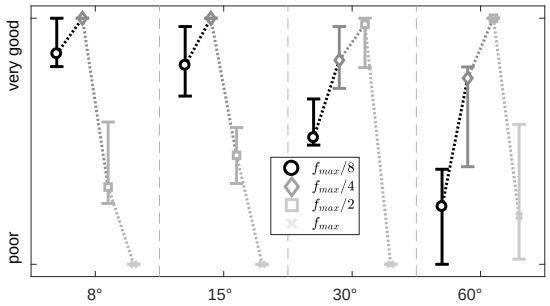
[1] https://puredata.info/

3.1 Pilot tests to determine optimal parameters

A large block size increases the frequency resolution and decreases the computational effort per sample, but it also increases latency and decreases the update rate, i.e. responsiveness to head-orientation changes. For instance, a block size of $N = 1024$ samples delays the update by more than 10 ms, which can reduce continuity and increase temporal artifacts at the overlapping synthesis blocks; especially for fast head movements. Our perceptual parameter optimization therefore considers smaller block sizes $N = \{128, 256, 512\}$, only.
The upper limit for the cross-over frequency f_c is defined by the grid resolution $\Delta\varphi$ which was $\{8°, 15°, 30°, 60°\}$ in our test.
In a simplistic model, the maximum delay between the adjacent signals becomes $\tau = \frac{r}{c}\sin(\Delta\phi)$, where $r = 8.5$ cm is the head radius and $c = 343$ m/s is the speed of sound. A first comb-filter notch is expected at $f_{max} = \frac{1}{2\tau} = \frac{c}{2r}\frac{1}{\sin(\Delta\phi)} \approx 2\,\text{kHz}\frac{57.3°}{[60°, 30°, 15°, 8°]} \approx [2, 4, 8, 16]\,\text{kHz}$ for $\Delta\phi = \{60°, 30°, 15°, 8°\}$. To find out how many octaves below to set the respective crossover, we test the frequencies $f_c(\Delta\phi) = 2^k f_{max}(\Delta\phi)$ with $k = \{-3, -2, -1, 0\}$.



(a) Perceived quality of binaural rendering for block sizes N of 128, 256 and 512 samples. The crossover frequencies f_c are set to $\{500, 1000, 2000, 4000\}$ Hz for the resolutions $\Delta\varphi$ of $\{8, 15, 30, 60\}^\circ$, respectively.



(b) Perceived quality of binaural rendering for crossover frequencies f_c which are either 1, 2, and 3 octaves below or at the analytical comb frequency f_{max} (first notch). The block size of the short-time processing is set to $N = 128$ samples.

Fig. 2: Median (markers) and IQR-based 95% confidence interval (solid lines) of ratings in the pilot test for all subjects. Rated was the perceived overall difference to the reference (linearly interpolated BRIRs on a 1° resolution).

To determine optimal block sizes and crossover frequencies from those candidates, two multi-stimulus pilot tests were performed 4 times by each of the 3 authors, using the same experimental environment, grid resolutions $\Delta\varphi$, and BRIRs as in the main experiment, cf. Sec. 4. Listeners were asked to rate the perceived quality, which is composed of spatial mapping, coloration, and temporal artifacts.

The first pilot test to find the optimal block size N used pink noise from the virtual center loudspeaker, and a cross-over frequency $f_c(\Delta\phi) = 2^{-2} f_{max}(\Delta\phi)$. As reference we used the linearly interpolated BRIR rendering with 1° resolution and a block size of 128 samples, as this setting allows for artifact free dynamic rendering, i.e. $f_{max}(1^\circ) \approx 120\,\text{kHz}$. The ratings of the perceived quality depending on the block size for varying grid resolutions are shown in Fig. 2(a). For all resolutions except 60°, a block size of 128 samples is significantly best ($p < 0.025$). In comparison, for 60°, the differences are smaller, revealing significant differences only between 128 and 512 samples ($p = 0.028$).

The second pilot test to determine the optimal crossover f_c used the optimal block size of $N = 128$ samples. Again, pink noise from the front was used as stimulus and the results in Figure 2(b) indicate that for 8° and 15° a cross-over at $f_{max}/4$ (two octaves below) is the optimal setting ($p < 0.024$), whereas the optimal crossover for 60° is 1 octave below f_{max} ($p < 0.008$). For 30°, $f_{max}/2$ is significantly better than f_{max} and $f_{max}/8$ ($p < 0.01$), but not significantly better than $f_{max}/4$ ($p = 0.22$).

Discussion: In all the tested grid resolutions of the first pilot test, a block size of $N = 128$ samples is preferable and offers the fastest response time.

Concerning the crossover: In the expected comb-filter response of an interpolated signal with $\alpha = 0.5$, a $3dB$ attenuation still appears if the phase offset between x_m and x_{m+1} is $\pi/2$. Therefore an artifact-free signal requires a crossover that is at least one, better two octaves lower than the comb-filter notch at $f_{max}(\Delta\varphi)$, which yields $f_c = [500, 1\,\text{k}, 2\,\text{k}, 4\,\text{k}]\,\text{Hz}$ for $\Delta\phi = [60^\circ, 30^\circ, 15^\circ, 8^\circ]$.

On the other hand, duplex theory [25, 26, 27] suggests that interaural time difference (ITD) is the dominant localization cue for frequencies below $1 \ldots 2\,\text{kHz}$, so two octaves may be too low for the coarse grids. While low crossover frequencies mitigate undesired switching and coloration artifacts, setting it too low counteracts successful localization in binaural synthesis.

The pilot study on the crossover suggests $f_c = [1, 2, 2, 4]\,\text{kHz}$ for the grids $\Delta\phi = [60^\circ, 30^\circ, 15^\circ, 8^\circ]$.

4 Listening Experiments

The final listening experiments consisted of two parts: (i) rating of spatialization and timbre compared to a reference condition for static rendering (four different head orientations), and (ii) rating of continuity or robustness when rendering dynamically, i.e. incorporating head movements of listeners.

We used the optimal settings as identified in the pilot tests and conducted further listening experiments

in order to compare the perceptual aspects of the proposed algorithm for different BRIR resolutions $\Delta\phi = \{8, 15, 30, 60\}^\circ$; linearly interpolated ear signals using the same resolutions were included in the experiment. The overall 10 test participants (all male and at an age of 27…42) were asked to rate the overall difference between a reference (artifact-free linear interpolation with 1° ($f_{max} = 120\,\text{kHz}$) resolution) and the test signals on a continuous scale from *poor* to *very good*. A hidden reference was used for screening of ratings, and thus the test procedure can be described as a multi-stimulus test on a continuous scale with a hidden reference, cf. MUSHRA [28]. The test signals were continuously looped and participants were allowed to seamlessly switch between signals in real-time as often as desired. The implementation of algorithm was done in *pure data* [2], an open source real-time audio software and for playback we used AKG K702 headphones equipped with the IEM headtracker [29]. Phase selection, see *Decider* in Fig. 1, used an update rate of 200 Hz.

The underling BRIRs were measured with a Neumann KU100 artificial head at the IEM production studio (volume $127\,\text{m}^3$, base area $42\,\text{m}^2$, $\text{T}_{60} \approx 0.4\,\text{s}$) equipped with Neumann KH310 A loudspeakers at various directions. The measurements were done with a resolution of 1° and using the exponentially swept sine technique (the deconvolved, and equalized BRIRs can be found here[3]).

For testing **timbre and spatialization** we evaluated four different static head orientations $\varphi = \{12, 21, 37, 78\}^\circ$ for a frontal source position and pink noise was used as source signal. Note that the orientations were chosen such that $0 < \alpha < 1$ (meaning interpolation is needed) for all tested resolutions and that the sign of the orientation was randomly changed across participants.

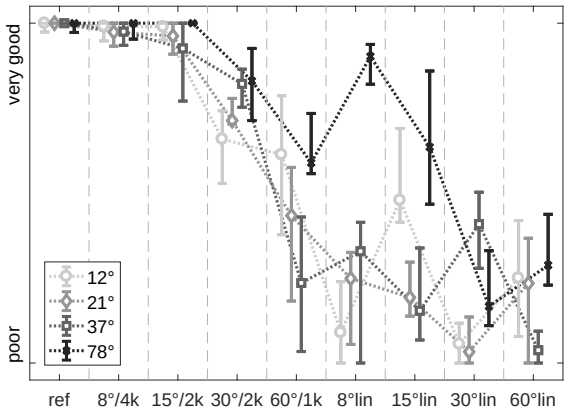
Testing the **continuity** involved pink noise and a music signal played back over a virtual frontal (0°) and lateral (90°) loudspeaker, respectively. Here listeners were asked to rotate their head between $\phi = -45^\circ \ldots 45^\circ$; a check-box for automatic rotation was included in the test to prevent neck pain of listeners.

While front-to-back confusion is implicitly evaluated via the lateral source, no dorsal source was included as no additional insight was expected (listeners are more sensitive to frontal positions).

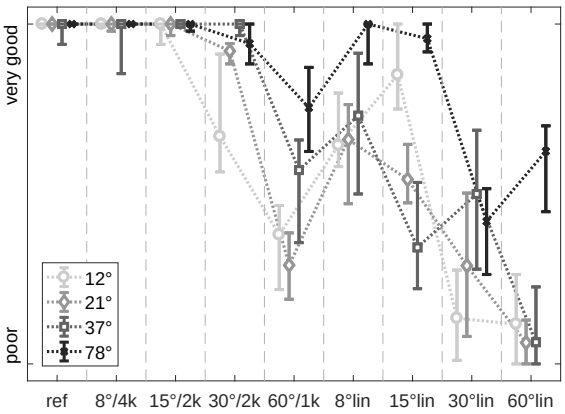
[2] https://puredata.info/

[3] https://opendata.iem.at/projects/binauralroomresponses/

The results of the experiments are depicted in Figs. 3 and 4 and are discussed in detail (per orientation and source signal) below. Significance of differences between median values was evaluated using a Kruskal-Wallis test. Note that when summarizing multiple comparisons only the p-value of the comparison with the least or most significant difference is given and indicated by $p < p_{max}$ and $p > p_{min}$, respectively.



(a) Timbre.



(b) Spatialization.

Fig. 3: Median (markers) and IQR-based 95% confidence interval (solid lines) of ratings from all 10 subjects for testing the perceived overall difference to the reference (linearly interpolated BRIRs on a 1° resolution and pink noise as test signal). Results are shown for a frontal source and head orientations of $\{12, 21, 37, 78\}^\circ$, respectively. Settings of the algorithm are indicated by $\Delta\varphi/f_c$, where *lin* denotes a full-bandwidth linear interpolation.

Timbre

78° *direction:* 8°/4k and 15°/2k are not significantly different from ref ($p > 0.16$). For the larger spacings, quality decreases significantly with the spacing ($p < 0.011$). For all resolutions, the proposed algorithm performs significantly better than linear interpolation ($p < 0.005$) and the best linear interpolation condition 8°lin is comparable to the 30°/2k ($p = 0.17$) condition.
37° *direction:* All conditions are significantly different from the reference, even 8°/4k ($p = 0.041$). Quality decreases with the spacing, however, 8°/4k is not significantly better than 15°/2k ($p = 0.21$) and 30°/2k not significantly worse than 15°/2k ($p = 0.088$). Except for 60°/1k ($p = 0.092$), the algorithm significantly increases the quality in comparison to linear interpolation ($p << 0.001$) and the best linear condition 30°lin is not significantly different from the worst dual-band approach 60°/1k ($p = 0.089$).
21° *direction:* All conditions are significantly different from the reference, even 8°/4k ($p = 0.036$). Quality decreases with the spacing, however, 8°/4k is not significantly better than 15°/2k ($p = 0.63$). The algorithm significantly increases the quality in comparison to linear interpolation ($p < 0.049$). Again the best linear condition is 8°lin is not significant different from the worst dual-band interpolation 60°/1k ($p = 0.096$).
12° *direction:* 8°/4k and 15°/2k are not significantly different from ref ($p > 0.41$). Quality decreases with the spacing, however, 30°/2k is not significantly better than 60°/1k ($p = 0.82$). The algorithm significantly increases the quality in comparison to linear interpolation ($p < 0.016$) and the best linear condition 15°lin is not significantly different from the worst dual-band settings 30°/2k and 60°/1k ($p > 0.25$).

Spatialization

78° *direction:* 8°/4k and 15°/2k are not significantly different from reference ($p > 0.19$). 60°/1k is significantly worse than the finer samplings ($p < 0.003$). Only for 30°/2k and 60°/1k, the quality significantly increases when applying the dual-band interpolation in comparison to linear interpolation ($p < 0.041$).
37° *direction:* all dual-band interpolation conditions are similar to the reference ($p > 0.35$), except 60°/1k ($p << 0.001$). For all resolutions, the dual-band approach significantly increases the quality in comparison to the full-band linear interpolation ($p < 0.003$) and the best linear condition 8°lin is not significantly different from the worst dual-band setting 60°/1k ($p = 0.076$).
21° *direction:* 8°/4k and 15°/2k are not significantly different from reference ($p > 0.25$). For all resolutions, the algorithm significantly increases the quality in comparison to linear interpolation ($p < 0.005$). Here, the best linear condition 8°lin and is significantly better than the worst dual-band setting 60°/1k ($p = 0.016$), but significantly worse than 30°/2k ($p << 0.001$).
12° *direction:* Only 8°/4k is not significantly different from ref ($p = 0.068$). Quality decreases with the spacing ($p < 0.03$) and the dual-band interpolation significantly increases the quality in comparison to linear interpolation ($p < 0.028$). The best linear condition 15°lin is significantly better than 60°/1k ($p = 0.002$) and not significantly different from 30°/2k ($p = 0.14$).

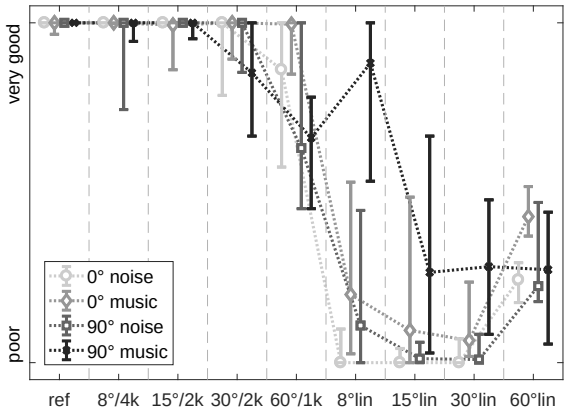


Fig. 4: Median (markers) and IQR-based 95% confidence interval (solid lines) of ratings from all 10 subjects for testing the perceived overall continuity or robustness in a dynamic rendering scenario for a frontal (0°) and lateral (90°) source. As reference linearly interpolated BRIRs on a 1° resolution were used. Settings of the algorithm are indicated by $\Delta\varphi/f_c$, where *lin* denotes a full-bandwidth linear interpolation.

Continuity

0° noise: From the dual-band implementations, only 60°/1k is significantly different from the reference ($p = 0.033$). The algorithm significantly increases the quality in comparison to linear interpolation ($p << 0.001$).
0° music: All dual-band settings are similar to the reference ($p > 0.41$) and significantly increase the quality in comparison to linear interpolation ($p < 0.008$). The best linear condition 60°lin is significantly better than 30°lin ($p = 0.001$).

90° noise: Among the dual-band settings only 60°/1k is significantly different from the reference ($p = 0.010$) and for all, a significant increase of quality in comparison to linear interpolation is found ($p < 0.004$). The best linear condition is 60°lin but it is not significantly better than 8°lin ($p = 0.225$).
90° music: Only 8°/4k and 15°/2k are not significantly different from the reference ($p > 0.35$). The dual-band interpolation significantly increases the quality in comparison to linear interpolation ($p < 0.006$), except for 8°lin ($p = 0.068$), which is also significantly better than all other linear conditions ($p < 0.034$).

5 Conclusion and Outlook

We presented a parameter study on the perceptually optimal settings of a dual-band real-time binaural room impulse response (BRIR) interpolation method which is used for dynamic binaural rendering. The low-frequency band is linearly interpolated in the time domain while the high-frequency band is spectrally interpolated in the short-time Fourier domain, where the magnitude is blended linearly and the phase for resynthesis is set to the phase of the BRIR which is nearest to the actual head orientation of the listener.
In a pilot test, the optimal settings of the block processing and cross-over frequencies were identified and later used in listening experiments which compared the proposed method for various underlying BRIR grid resolutions against full-bandwidth linear interpolation. The tested perceptual attributes included (i) timbre and spatialization, which were evaluated for four static orientations, and (ii) the continuity or robustness when rendering dynamically, i.e. real-time head rotations. We found that the dual-band approach significantly improves rendering quality when compared to the corresponding linear interpolation and suggest using a grid resolution of 15° and a cross-over frequency of 2 kHz as these settings unite efficiency and rendering quality. Demos and an implementation of the algorithm can be found at `https://opendata.iem.at/projects/binauralroomresponses/`.
As measuring BRIRs for various orientations is not hardware efficient and modular, an approach similar to [30] can be used for efficient measurements and individualization of BRIRs.

References

[1] Møller, H., "Fundamentals of binaural technology," *Applied Acoustics*, 36(3-4), pp. 171–218, 1992, doi:10.1016/0003-682X(92)90046-U.

[2] Pollack, I. and Trittipoe, W., "Binaural listening and interaural noise cross correlation," *J. Acoust. Soc. Am.*, 31(9), pp. 1250–1252, 1959, doi:10.1121/1.1907852.

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

[4] Plenge, G., "On the problem of 'in head localization'," *Acta Acustica united with Acustica*, 26(5), pp. 241–252, 1972, doi:10.1007/BF00026991.

[5] Werner, S., Klein, F., Mayenfels, T., and Brandenburg, K., "A summary on acoustic room divergence and its effect on externalization of auditory events," *8th International Conference on Quality of Multimedia Experience*, 2016, doi:10.1109/QoMEX.2016.7498973.

[6] Cubick, J., *Investigating distance perception, externalization and speech intelligibility in complex acoustic environments Hearing*, Ph.D. thesis, Technical University ofDenmark, 2017.

[7] Lindau, A., *Binaural Resynthesis of Acoustical Environments - Technology and Perceptual Evaluation*, Ph.D. thesis, 2014.

[8] Smyth, S. M., "Personalized headphone virtualization," *US Patent 7,936,887*, 2011.

[9] Smyth, S., Smyth, M., and Cheung, S., "Smyth SVS Headphone Surround Monitoring for Studios," *23rd UK Conference of the Audio Engineering Society*, pp. 1–7, 2008.

[10] Satongar, D., Lam, Y. W., and Pike, C., "Measurement and Analysis of a Spatially Sampled Binaural Room Impulse Response Dataset," in *21st International Congress on Sound and Vibration*, pp. 1–8, Beijing, 2014, ISBN 9781634392389.

[11] Begault, D. R., Wenzel, E. M., and Anderson, M. R., "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial

Perception of a Virtual Speech Source," *J. Audio Eng. Soc.*, 49(10), pp. 904–916, 2001.

[12] Brungart, D. S., Kordik, A. J., and Simpson, B. D., "Effects of headtracker latency in virtual audio displays," *Journal of the Audio Engineering Society*, 54(1-2), pp. 32–44, 2006.

[13] Hendrickx, E., Stitt, P., Messonnier, J.-C., Lyzwa, J.-M., Katz, B. F., and de Boishéraud, C., "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis," *J. Acoust. Soc. Am.*, 141(3), pp. 2011–2023, 2017, doi:10.1121/1.4978612.

[14] Engdegard, J., Resch, B., Falch, C., Hellmuth, O., Hilpert, J., Hoelzer, A., Breebaart, J., Koppens, J., Schuijers, E., and Oomen, W., "Spatial Audio Object Coding (SAOC) - The Upcoming MPEG Standard on Parametric Object Based Audio Coding," *124th AES Convention*, 2008.

[15] Jot, J.-M., Larcher, V., and Pernaux, J.-M., "A Comparative Study of 3-D Audio Encoding and Rendering Techniques," in *AES 16th International Conference on Spatial Sound Reproduction*, 1999.

[16] Pinchon, D. and Hoggan, P. E., "Rotation matrices for real spherical harmonics: General rotations of atomic orbitals in space-fixed axes," *Journal of Physics A: Mathematical and Theoretical*, 40(7), pp. 1597–1610, 2007, doi:10.1088/1751-8113/40/7/011.

[17] Lindau, A., "The Perception of System Latency in Dynamic Binaural Synthesis," *Fortschritte der Akustik: Tagungsband der 35. DAGA*, (1), pp. 1063–1066, 2009.

[18] Stitt, P., Hendrickx, E., Messonnier, J.-C., and Katz, B., "The Influence of Head Tracking Latency on Binaural Rendering in Simple and Complex Sound Scenes," *Audio Engineering Society Convention Paper*, pp. 1–8, 2016.

[19] Lindau, A., Maempel, H.-J., and Weinzierl, S., "Minimum BRIR grid resolution for dynamic binaural synthesis," *J. Acoust. Soc. Am.*, 123(5), p. 3498, 2008, doi:10.1121/1.2934364.

[20] Algazi, V. R., Duda, R. O., and Thompson, D. M., "Motion-tracked binaural sound," *AES: Journal of the Audio Engineering Society*, 52(11), pp. 1142–1156, 2004.

[21] Pruša, Z., Balazs, P., and Søndergaard, P. L., "A Non-iterative Method for STFT Phase (Re)Construction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(5), pp. 1154–1164, 2016.

[22] Hom, R. C.-M., Algazi, V. R., and Duda, R. O., "High-Frequency Interpolation for Motion-Tracked Binaural Sound," *Proceedings of the 121st AES Convention*, 2006.

[23] Lindau, A. and Roos, S., "Perceptual evaluation of discretization and interpolation for motion-tracked binaural (MTB) recordings," *26th Tonmeistertagung*, (November), pp. 680–701, 2010.

[24] Zhu, X., Beauregard, G. T., and Wyse, L. L., "Real-time signal estimation from modified short-time fourier transform magnitude spectra," *IEEE Transactions on Audio, Speech and Language Processing*, 15(5), pp. 1645–1653, 2007, doi:10.1109/TASL.2007.899236.

[25] Rayleigh, L., "On our perception of sound direction," *Philosophical Magazine Series 6*, 13(74), pp. 214–232, 1907, doi:10.1080/14786440709463595.

[26] Macpherson, E. A. and Middlebrooks, J. C., "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *J. Acoust. Soc. Am.*, 111(5), pp. 2219–2236, 2002, doi:10.1121/1.1471898.

[27] Hartmann, W. M., Rakerd, B., Crawford, Z. D., and Zhang, P. X., "Transaural experiments and a revised duplex theory for the localization of low-frequency tones," *J. Acoust. Soc. Am.*, 139(2), pp. 968–985, 2016, doi:10.1121/1.4941915.

[28] International Telecommunication Union, "ITU-R BS.1534-3, Method for the subjective assessment of intermediate quality level of audio systems," *ITU-R Recommendation*, 1534-3, pp. 1534–3, 2015.

[29] Romanov, M., Berghold, P., Rudrich, D., Zaunschirm, M., Frank, M., and Zotter, F., "Implementation and evaluation of a low-cost head-tracker for binaural synthesis," in *142nd Audio Engineering Society International Convention 2017, AES 2017*, 2017.

[30] Zaunschirm, M., Frank, M., and Zotter, F., "BRIR Synthesis Using First-Order Microphone Arrays," in *Audio Engineering Society Conference 144*, pp. 1–10, 2018.

5.3 Binaural Rendering with Measured Room Responses: First-Order Ambisonic Microphone vs. Dummy Head

This work was published as:

M. Zaunschirm, M. Frank, and F. Zotter. (2020). Binaural Rendering with Measured Room Responses: First-Order Ambisonic Microphone vs. Dummy Head. *Applied Sciences*, 10(5),

The idea and concept of this article were outlined by all authors. I, as first author wrote the original draft of the manuscript with help from the third author and periodical contributions from the second author. The revision and editing was done by me and the third author. I did most of the programming and graphical work, and prepared the samples for the listening experiment. The listening experiment was programmed by the third author and me with periodic contributions from the second author. The listening experiment was designed by all authors equally and the data was analyzed by the second author and me. All underlying measurements were done by the second author and myself.

applied sciences

MDPI

Article

Binaural Rendering with Measured Room Responses: First-Order Ambisonic Microphone vs. Dummy Head

Markus Zaunschirm *, Matthias Frank and Franz Zotter

Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Inffeldgasse 10/III, 8010 Graz, Austria; frank@iem.at (M.F.); zotter@iem.at (F.Z.)

* Correspondence: zaunschirm@iem.at

Received: 11 January 2020; Accepted: 21 February 2020; Published: 29 February 2020

Abstract: To improve the limited degree of immersion of static binaural rendering for headphones, an increased measurement effort to obtain multiple-orientation binaural room impulse responses (MOBRIRs) is reasonable and enables dynamic variable-orientation rendering. We investigate the perceptual characteristics of dynamic rendering from MOBRIRs and test for the required angular resolution. Our first listening experiment shows that a resolution between 15° and 30° is sufficient to accomplish binaural rendering of high quality, regarding timbre, spatial mapping, and continuity. A more versatile alternative considers the separation of the room-dependent (RIR) from the listener-dependent head-related (HRIR) parts, and an efficient implementation thereof involves the measurement of a first-order Ambisonic RIR (ARIR) with a tetrahedral microphone. A resolution-enhanced ARIR can be obtained by an Ambisonic spatial decomposition method (ASDM) utilizing instantaneous direction of arrival estimation. ASDM permits dynamic rendering in higher-order Ambisonics, with the flexibility to render either using dummy-head or individualized HRIRs. Our comparative second listening experiment shows that 5th-order ASDM outperforms the MOBRIR rendering with resolutions coarser than 30° for all tested perceptual aspects. Both listening experiments are based on BRIRs and ARIRs measured in a studio environment.

Keywords: binaural synthesis; dynamic binaural rendering; BRIR measurements; head-tracked binaural; psychoacoustics

1. Introduction

Typically, binaural rendering involves a convolution of source signals with measured or modeled head-related impulse responses (HRIRs) or binaural room impulse responses (BRIRs) and playback of the corresponding ear signals via headphones [1]. Both HRIRs and BRIRs implicitly contain the cues accessible to the human auditory system to perceive sound from a certain direction and distance, with a certain source width, envelopment, or spaciousness, cf. [2,3]. In order to reduce poor externalization when using both individual and non-individual HRIRs, it is often helpful to involve a natural or simulated acoustic room and thus to render with BRIRs instead. It is shown in [1,4] that BRIRs measured with a loudspeaker and a dummy head can achieve static binaural rendering of high audio quality and of convincing realism. Such a BRIR-based virtualization of loudspeakers in rooms is not only useful to virtualize multi-channel loudspeaker setups in mixing studios [5–7], but also to document and preserve acousmatic or electroacoustic music sceneries.

By involving the natural interaction of the ear signals with the head rotation of a listener, i.e., head-tracking [8] and dynamic rendering, immersion is improved as this reduces localization ambiguities and poor externalization [9–11]. Dynamic and interactive head-tracked BRIR rendering requires the acquisition of MOBRIRs (multi-orientation BRIRs), which can be tedious for highly resolved orientations. For best individualized results, in particular, one would need to measure

MOBRIRs of each individual listener in each room to be auralized. Efficient and versatile alternatives [12–14] propose to measure the listener-dependent (HRIRs) and room-dependent (RIRs) parts separately to enable individualization as a second step.

State of the art: Linear interpolation of coarse-orientation MOBRIRs can cause strong comb-filter artifacts. Lindau [15] showed for multiple-orientation binaural recordings that the minimum required binaural grid resolution to avoid artifacts is most sensitive in anechoic conditions, and less sensitive reverberant cases, in which particular reverberation did not matter. To ensure continuous and robust interpolation from orientations coarser than 3°, a dual-band interpolation strategy is required, which literature refers to as motion-tracked binaural (MTB) [16]. At low frequencies, the dual-band approach interpolates the headphone signal from neighboring pairs of recorded ear signals linearly, while comb filters at high frequencies are avoided by combining interpolated spectral magnitudes with suitable phase values, e.g., found by spectrogram inversion [17]. Less challenging approaches yielding a suitable phase are discussed in [18] and perceptual properties are studied in [19]. Perceptually optimal cross-over frequencies and block sizes were investigated in [20] for the static and dynamic case, with which rendering from MOBRIRs resolved finer than 30° was found to be indiscernible from a 1°-resolved reference.

Dynamic rendering based on first-order Ambisonic RIRs (ARIRs) and a pre-measured set of high-resolution HRIRs, e.g., of a dummy head [21], is studied in [22]. Static rendering with ASDM (Ambisonic Spatial Decomposition Method) upscaling was shown to yield perceptually indistinguishable results for 7th order when compared to a reference dummy-head BRIR. Moreover, the study involved three rooms of different reverberation times (0.3 s, 0.7 s, and 1.4 s) and could show that the performance of the ASDM method did not depend on the particular reverberation time.

Contents: ARIR-based and MOBRIR-based rendering haven't been compared yet, and our contribution deals with establishing a balanced comparison. The goal is to find out configurations in which both methods yield perceptually optimal or correspondingly scaled results. Some correspondence is expected, as both the binaural Ambisonic rendering of ARIRs using MagLS [22,23] and the interference-avoiding high-frequency strategy of MOBRIR-based rendering [20] rely on spectral phase simplification at high frequencies, and both relate to an angular resolution. To make the comparison reproducible, a room impulse response data set (dummy head and Ambisonic microphone) is measured in a studio environment and made accessible in this contribution. As the main part of the paper, Section 3 is dedicated to our comparative listening experiment on variable-orientation rendering from MOBRIRs resolved in {30°, 45°, 60°} steps and corresponding ASDM-based Ambisonic orders {5, 3, 1} rendered using the same dummy head HRIRs [21].

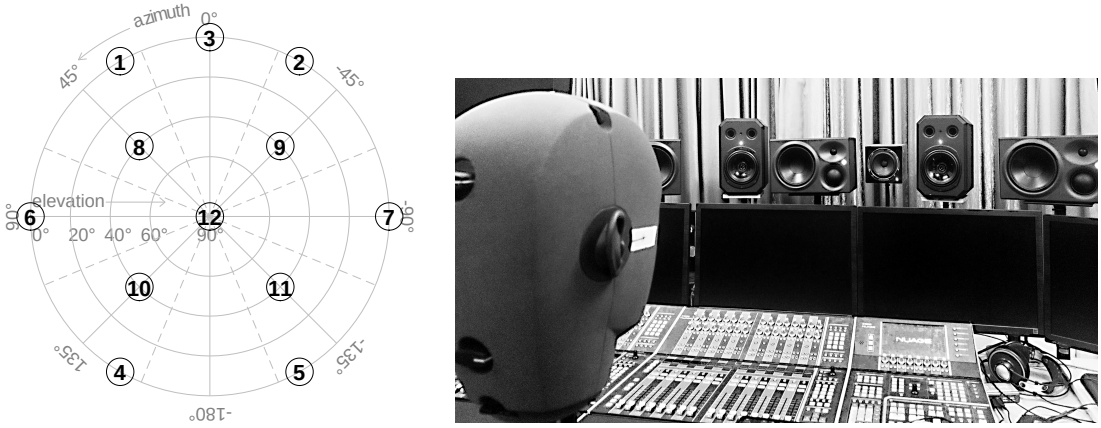
As the Ambisonic renderer is already available (https://plugins.iem.at/), we dedicate Section 2 of the paper to supporting open research into MOBRIR-based rendering by providing an implementation example (Appendix A), example renderings (https://phaidra.kug.ac.at/view/o:77319), listening test response data (https://opendata.iem.at/projects/listening_experiment_data/), and a summarized statistical analysis of research presented at AES IIA [20].

In both listening experiments, participants are asked to rate for static rendering the perceptual attributes (i) timbre, (ii) spatial mapping, and for dynamic rendering (iii) its continuity. Both experiments compare the renderers to coarse linear MOBRIR interpolation (anchor) and to linear interpolation from 1° MOBRIRs (reference).

Measurements: The RIR measurements used here are available online (https://opendata.iem.at/projects/binauralroomresponses/) and were taken from the IEM production studio (volume 127 m^3, base area 42 m^2, $T_{60} \approx 0.4$ s) in which Neumann (Germany) KH310-A loudspeakers are mounted in various directions. MOBRIRs were measured with a Neumann KU100 dummy head in rotations of 1° steps (turntable) using the exponentially swept sine technique. Available loudspeaker directions that are depicted in Figure 1a, 1b show the dummy head in the center listening position, facing the center loudspeaker (channel 3). The B-format ARIRs were measured after replacing turntable and dummy head with the Soundfield ST450 array. The room was selected for studying MOBRIR

Appl. Sci. **2020**, *10*, 1631

interpolation and binaural Ambisonic rendering as the presence of its short reverberation already supports externalization in typical listening environments and its pronounced direct and early parts are expected to be critical considering both timbral artifacts and spatial mapping deficiencies [15,24].



(**a**) Loudspeaker layout of the IEM production studio.

(**b**) KU100 dummy head in the center of the loudspeaker setup.

Figure 1. Binaural room impulse response measurement setup.

2. Experiment I: Dynamic Rendering of Multiple-Orientation Binaural Signals

Experiment I is based on data of a previous study [20], and it evaluates the dual-band strategy *linear interpolation with switched high-frequency phase* (LISHPh) using dummy-head MOBRIRs of the resolutions 8°, 15°, 30°, and 60°, and compares it with a reference linear interpolation of MOBRIRs resolved by 1°. The LISHPh method is described in Section 2.1, and the design and implementation of the listening experiment are discussed in Sections 2.2 and 2.3, respectively. Response data[3] and examples of the audio stimuli[2] are made available for download; examples include renderings of static head-orientations and of emulated continuous head rotation. Finally, the results of *Experiment I* are discussed in Section 2.4.

2.1. Linear Interpolation with Switched High-Frequency Phase (LISHPh)

For both the left and right ear, the interpolated ear signal in a horizontal set of orientations is obtained by a combination of the corresponding signals $x_q(t)$ and $x_{q+1}(t)$ belonging to the head orientation closest to the current orientation of the listener $\varphi(t)$, where t is the discrete-time index. With MOBRIRs measured for Q equi-angular orientations on the horizon (around the Cartesian z-axis), and $\Delta\varphi$ as azimuthal resolution, the indices of the two closest BRIRs (or recorded ear signals) are $q = \left\lfloor \frac{\varphi(t)}{\Delta\varphi} \right\rfloor$, and $q+1 = \left\lceil \frac{\varphi(t)}{\Delta\varphi} \right\rceil$, where $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ are the floor and ceil functions, respectively, and the ear signals $x_q(t)$, and $x_{q+1}(t)$ are obtained by convolution; see Figure 2a.

In a broadband linear interpolation, the resulting ear signal is obtained as

$$x(t) = (1-\alpha)x_q(t) + \alpha x_{q+1}(t), \tag{1}$$

where the interpolation weight is obtained by $\alpha = \left\lceil \frac{\varphi(t)}{\Delta\varphi} \right\rceil - \frac{\varphi(t)}{\Delta\varphi}$. However, the linear combination of delayed signals produces comb-filtering introducing severe colorations in the resulting signal. In particular, the maximum delay $\tau = \frac{r}{c}\sin(\Delta\phi)$ between adjacent HRIRs is estimated by a simplistic head model, where $r = 8.5\,\text{cm}$ is the head radius and $c = 343\,\text{m/s}$ is the speed of sound. The maximum delay is observed between ear signals of the head orientation 0° and those of the orientations $\pm\Delta\varphi$,

Appl. Sci. **2020**, *10*, 1631

for a frontal source. To avoid destructive interference, artifact-free linear interpolation can only be achieved below

$$f_{max} = \frac{1}{2\tau} = \frac{c}{2r}\frac{1}{\sin(\Delta\phi)}. \tag{2}$$

Spectral artifacts of linearly interpolated BRIRs are comparable with those of HRIRs when the direct sound dominates. For interpolation of BRIRs from a diffuse field, the same (worst-case) frequency limit for destructive interference holds. In particular, if the contribution of frontal sounds is pronounced, the interpolated result partly contains the destructive interference at f_{max}.

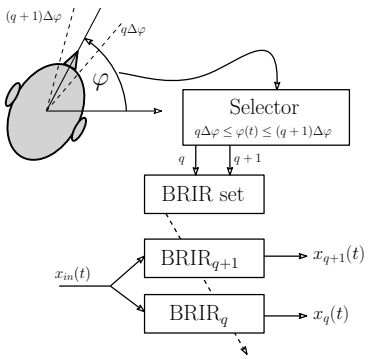
The LISHPh method is employed [16,18,20] to avoid noticeable spectral artifacts around and above f_{max}, regardless of the acoustic scenario. As depicted in Figure 2b, the signal in the lower band is processed in the time domain by applying the linear weights as in Equation (1), and the signal in the high-frequency band is obtained by magnitude interpolation

$$X(k) = \{(1-\alpha)|X_q(k)| + \alpha|X_{q+1}(k)|\}\, \mathrm{e}^{i\measuredangle(k)}, \tag{3}$$

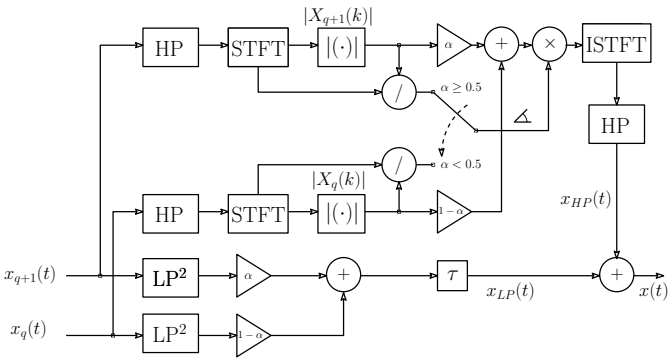
where k is the frequency index of a short-time Fourier transform (STFT) frame, $i^2 = -1$, and $\measuredangle(k)$ is the phase argument which is switched between $\measuredangle(k) = \measuredangle_{q+1}(k)$ for $\alpha \geq 0.5$ and $\measuredangle(k) = \measuredangle_q(k)$ else. Whenever the phase argument of a narrow-band signals has to make a transition by π, switching can theoretically become audible and can only be avoided by spectrogram inversion [17,18]. We avoided the additional effort, as the negative influence of the switching noise turned out to be inaudible for speech, music recordings, noise, etc. with suitable block size settings [20].

2.2. Listening Experiment: Design

We tested the LISHPh method rendering from MOBRIR resolutions of $\Delta\phi = \{8, 15, 30, 60\}^\circ$ and also included the broadband linearly interpolated ear signals for comparison. During the listening experiment, each listener was asked to (i) rate the spatial mapping (i.e., direction and distance) and timbre compared to a reference condition for static rendering (four different head orientations), and (ii) to rate the continuity or robustness when rendering dynamically, i.e., incorporating head movements of the listener.



(a) Filter-switching and convolution to obtain the two neighbouring ear signals which are closest to the current head rotation.



(b) Block diagram of the LISHPh interpolation method. The lower band is interpolated linearly in the time domain, while the high-frequency band is interpolated in the short-time Fourier domain. The high-frequency signal is obtained by a magnitude interpolation and the phase for reconstruction is chosen dependent on α (interpolation weight).

Figure 2. Block diagram for BRIR selection, convolution, and continuous interpolation for one ear in a dynamic binaural rendering scenario.

The 10 test participants (all male and at an age between 27 and 42) were asked to rate the overall difference between a reference (artifact-free linear interpolation with 1° resolution) and the

Appl. Sci. **2020**, *10*, 1631

test signals on a continuous scale from *poor* to *very good*. A hidden reference was used for screening of ratings, and thus the test procedure can be described as MUSHRA-like (multi stimulus with hidden reference and anchor [25]). The test signals were continuously looped, and participants were allowed to seamlessly switch between signals in real-time as often as desired.

In terms of **timbre and spatialization**, we tested four different static head orientations $\varphi = \{12, 21, 37, 78\}^\circ$ (the sign of the orientation was randomly changed across participants) for a frontal source position with pink noise and music as source signal, respectively. The choice of the source position and orientations was met to make the experiment most sensitive to the expected interpolation artifacts: on the one hand, time-delays (phases or ITDs) are most head-orientation-dependent for predominantly frontal source positions, and, on the other hand, orientations were selected to enforce interpolation with $0.2 < \alpha < 0.8$ for all MOBRIR sets under test.

Testing the **continuity** involved a pink noise and a music signal played back over a virtual frontal (0°) and lateral (90°) loudspeaker. Here, listeners were asked to rotate their head between $\phi = -45^\circ \ldots 45^\circ$ and a check-box for automatic rotation with 180°/1s was included for fast movements.

2.3. Listening Experiment: Implementation and Settings

The real-time implementation of the LISHPh, as well as the broadband linear interpolation, was done in *pure data* (https://puredata.info/), an open source real-time audio software. Appendix A describes the example implementation provided online (https://phaidra.kug.ac.at/o:97087).

Block processing: In the short-time block processing with block size N and hop size $L = N/2$ of the high-frequency part, a sine half-wave window is applied at both the analysis and synthesis stage to reduce cyclic artifacts at the block boundaries. As found in [20] and as the optimum for broadband musical sounds, it is crucial to keep the block size low to avoid temporal artifacts and to obtain low latency. Thus, we suggest setting $N = 128$ at a sampling rate of 44.1 kHz. For high-frequency phase selection, we used an update rate of 200 Hz; see the block labeled as *Selector* in the diagram of Figure 2a.

Crossover: We estimate the spectral ripple of two interfering signals with phase offsets by $\frac{1}{2}[e^{\mathrm{i}\frac{\phi}{2}} + e^{-\mathrm{i}\frac{\phi}{2}}] = \cos\frac{\phi}{2}$. To keep the spectral ripple below 3 dB, we require a phase difference $\phi < \frac{\pi}{2}$, or $\phi < \frac{\pi}{4}$ to keep it below 0.7 dB, and hence spectrally inaudible [26]. With Equation (2) or a rule of thumb $f_{max} \approx 2\,\mathrm{kHz}\frac{57.3^\circ}{\Delta\varphi}$, the phase difference is $\phi = \frac{\pi f}{f_{max}}$ in our case. Spectrally, good results are achieved with a crossover frequency $f_c \approx \frac{f_{max}}{4}$. However, setting it too low, e.g., $f_c < 1.5$kHz, impedes interaural phase cues in a relevant frequency range. Accordingly, the choice

$$f_c(\Delta\phi) = 2^u\, f_{max}(\Delta\phi), \tag{4}$$
$$\text{with} \quad u = \begin{cases} -2, \text{ if } \Delta\varphi < 30^\circ \\ -1, \text{ if } \Delta\varphi \geq 30^\circ \end{cases}$$

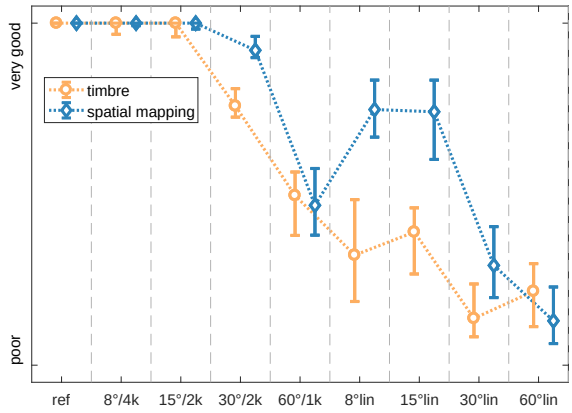
offers a reasonable trade-off, cf. [20]. For $\Delta\phi = \{8, 15, 30, 60\}^\circ$, we obtain the crossovers at $f_c = [4, 2, 2, 1]$ kHz. The crossover is implemented by 4th-order Linkwitz–Riley filters because of their in-phase sub bands [27,28].

For playback, we used AKG K702 headphones equipped with the IEM headtracker [8] and the experiment was conducted in a quiet office room.

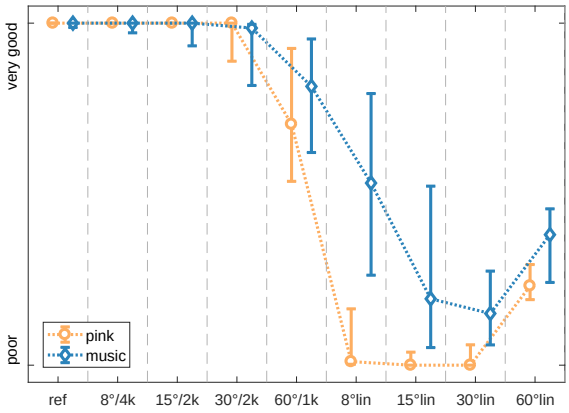
2.4. Results and Discussion

The statistical analysis below uses pooled data for each attribute; see Figure 3. For timbre and spatial mapping, ratings of four directions were pooled, and both virtual loudspeaker directions were pooled for continuity. Please note that throughout the article we use a Wilcoxon signed-rank test [29] with a Bonferroni–Holm correction [30] to determine p-values of pair-wise comparisons between test conditions and define $p < 0.05$ as significantly different. We employ non-parametric statistics as we do not assume a normal distribution of the ratings due to severe clustering at the limits of the scale.

Appl. Sci. **2020**, *10*, 1631



(a) Timbre and spatial mapping of the pooled data for all tested directions for a frontal source and head orientations of $\varphi = \{12, 21, 37, 78\}^\circ$.



(b) Continuity of the pooled data (virtual speaker at front and side, respectively).

Figure 3. Median (markers) and 95% confidence interval (solid lines) of ratings from all 10 subjects for testing the perceived difference to the reference (linearly interpolated BRIRs on a 1° resolution). Settings of the algorithm are indicated by $\Delta\varphi/f_c$, where *lin* denotes a broadband linear interpolation.

Timbre: Per MOBRIR resolution, there is a clear advantage of the LISHPh method in the settings proposed compared to broadband linear interpolation. The p-values (significance level) given in the upper triangle of Table 1 indicate that there are four groups, which are significantly different from each other. The LISHPh interpolations with settings 8°/4k and 15°/2k are not significantly different ($p = 0.11$) to the 1° reference condition and perform significantly better than all other conditions. For the coarser resolutions, the quality of the LISHPh interpolation decreases significantly with spacing. However, for all orientation resolutions, LISHPh performs significantly better than linear interpolation ($p < 0.005$). The best linear interpolation conditions 8°lin and 15°lin are comparable to the worst LISHPh condition 60°/1k ($p > 0.36$).

Table 1. p-values (Wilcoxon signed-rank test with Bonferroni–Holm correction) for ratings of timbre and spatial mapping of Experiment I. The upper triangle corresponds to timbre, the lower triangle to spatial mapping. Insignificant differences (p-values ≥ 0.05) are indicated by bold numbers.

Method	ref	**8°/4k**	**15°/2k**	**30°/2k**	**60°/1k**	**8°lin**	**15°lin**	**30°lin**	**60°lin**
ref	-	**0.11**	**0.11**	0.00	0.00	0.00	0.00	0.00	0.00
8°/4k	**0.13**	-	**0.95**	0.00	0.00	0.0	0.00	0.00	0.00
15°/2k	**0.58**	**0.35**	-	0.00	0.00	0.00	0.00	0.00	0 .00
30°/2k	0.00	0.02	0.00	-	0.00	0.00	0.00	0.00	0.00
60°/1k	0.00	0.00	0.00	0.00	-	**0.72**	**0.36**	0.00	0.00
8°lin	0.00	0.00	0.00	0.00	0.00	-	**0.95**	0.03	**0.16**
15°lin	0.00	0.00	0.00	0.01	0.01	**0.58**	-	0.00	0.00
30°lin	0.00	0.00	0.00	0.00	0.02	0.00	0.00	-	**0.95**
60°lin	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.13**	-

Spatial Mapping: The results for spatial mapping, i.e., localization and distance impression, indicate that there is no significant difference between the ref and 8°/4k, and 15°/2k ($p > 0.13$) conditions, cf. lower triangle in Table 1. For coarser resolutions, the quality of spatial mapping decreases significantly. Again, the LISHPh conditions clearly outperform the linear interpolation. However, the best linearly interpolated conditions 8°lin and 15°lin are significantly better than the worst LISHPh condition

60°/1k. This is caused by its low crossover frequency at $f_c = 1$kHz which is set to avoid comb filtering (cf. Equation (5)) but already distorts the interaural time difference (ITD) in a sensitive frequency range [31].

Continuity: The results for perceived continuity are depicted in Figure 3b. While there is clear absolute difference depending on the source signal (yellow vs. blue line-pink noise vs. music), the trend is similar. For both signal types, all of the LISHPh conditions except 60°/1k do not significantly differ from the reference and from each other; see Table 2. Coarser MOBRIR resolutions lead to a decrease in quality, and LISHPh conditions clearly outperform linear interpolation of corresponding resolution. The improved continuity for 60°lin compared to the denser MOBRIRs can be explained by its reduced timbral variation when rotating the head, which seemed more important to listeners than spatial mapping.

Table 2. *p*-values (Wilcoxon signed-rank test with Bonferroni–Holm correction) for ratings of continuity in Experiment I. The upper triangle corresponds to the pink noise, the lower triangle to music as source signal. Insignificant differences (p-values ≥ 0.05) are indicated by bold numbers.

method	ref	**8°/4k**	**15°/2k**	**30°/2k**	**60°/1k**	**8°lin**	**15°lin**	**30°lin**	**60°lin**
ref	-	**2.19**	**1.69**	**0.90**	0.02	0.00	0.00	0.00	0.00
8°/4k	**1.00**	-	**1.78**	**1.69**	**0.15**	0.0	0.00	0.00	0.00
15°/2k	**1.06**	**1.70**	-	**0.44**	0.00	0.00	0.00	0.00	0 .00
30°/2k	**0.86**	**0.82**	**1.33**	-	0.02	0.00	0.00	0.00	0.00
60°/1k	**0.09**	**0.15**	**0.90**	**0.86**	-	0.00	0.00	0.00	0.00
8°lin	0.01	0.03	0.02	**0.09**	**0.83**	-	**0.16**	**1.38**	**0.39**
15°lin	0.01	0.00	0.02	0.03	**0.09**	**0.90**	-	**2.19**	0.00
30°lin	0.00	0.00	0.00	0.00	0.00	0.02	**0.88**	-	0.00
60°lin	0.00	0.00	0.01	0.00	0.00	**1.06**	**1.70**	**0.33**	-

3. Experiment II: Dummy-Head MOBRIR vs. ARIR

In *Experiment II*, we evaluate and compare the perceptual aspects of MOBRIR (LISHPh) and ARIR-based dynamic rendering (ASDM). The concept of ARIR-based rendering and the relevant signal processing involved to accomplish upscaling are described in Section 3.1. A description of the listening experiment, the implementation, and the corresponding discussions are presented in Section 3.2, 3.3, and 3.4, respectively. Appendix B shows the *MATLAB* implementation of the ASDM upscaling. Audio examples[2] of the material used in the listening experiment as well as its response data [3] are available for download. The examples include renderings of static head-orientations and of emulated continuous head-orientations.

3.1. Rendering with Measured Ambisonic RIR and the Ambisonic Spatial Decomposition Method (ASDM)

As depicted in Figure 4, dynamic binaural rendering from room responses measured in Ambisonics (ARIRs) is modular and consists of three blocks: (i) a multi-channel convolution of the source signal with an order N upscaled ARIR, (ii) an efficient rotation [32,33] corresponding to the head orientation of the listener, (iii) and multi-channel convolution with an Ambisonic binaural renderer.

For efficient and low-effort measurements of the ARIR, we use a compact first-order tetrahedral spherical microphone array and denote the discrete-time B-format ARIRs $h(t)$, $x(t)$, $y(t)$, $z(t)$ as the responses of a Soundfield ST450 array.

Appl. Sci. **2020**, *10*, 1631

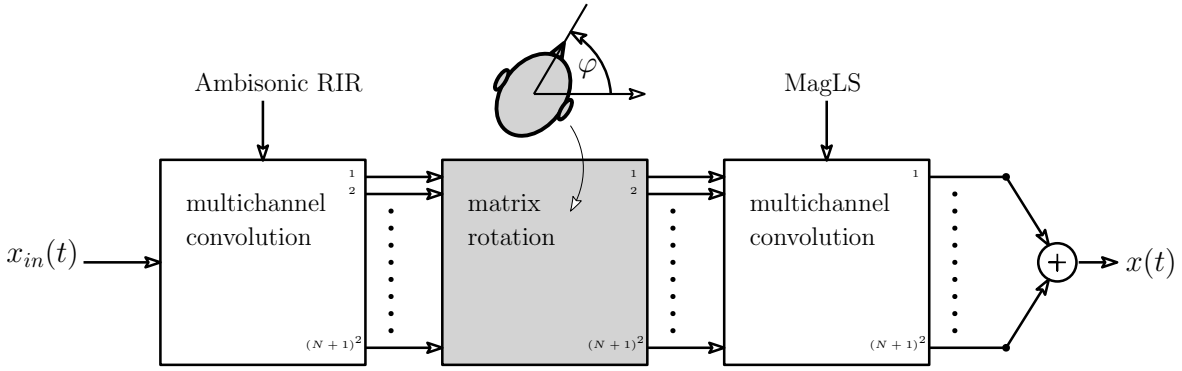


Figure 4. Block diagram of binaural rendering with measured Ambisonic RIRs (ARIRs). Here, *MagLS* refers to a state-of-the art Ambisonic binaural render.

Similar to the spatial decomposition method SDM [34], the Ambisonic SDM (ASDM) assigns a direction of arrival (DOA) $\boldsymbol{\theta}(t)$ to each discrete-time sample t of the omni-directional RIR component $h(t)$, cf. [22]. For the DOA estimation, we suggest using the pseudo-intensity vector (PIV) [35] for the frequencies between 200 Hz and 3 kHz. Here, the upper frequency limit is chosen below the spatial aliasing frequency $f_a = \frac{c}{2\pi r_{ST450}} \approx 3.6$kHz for $r_{ST450} = 1.5$cm and the low cut minimizes low-frequency disturbance in the DOA estimation. We perform a zero-phase band limitation (e.g., by MATLAB's filtfilt) denoted by $F_{200-3\mathrm{k}}$ and a zero-phase temporal smoothing F_L of the resulting PIV using a moving-average Hann window in the interval $[-L/2; L/2]$ for $L = 16$ to get the DOA estimate

$$\boldsymbol{\theta}(t) = \frac{\tilde{\boldsymbol{\theta}}(t)}{\|\tilde{\boldsymbol{\theta}}(t)\|}, \quad \text{with} \tag{5}$$

$$\tilde{\boldsymbol{\theta}}(t) = F_L \left\{ F_{200-3k}\{h(t)\}\; F_{200-3k} \left\{ \begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix} \right\} \right\}$$

as Cartesian unit vector $\boldsymbol{\theta}(t)$.

In a first step, the ASDM-upscaled ARIR re-encodes every time sample at the detected DOA

$$\tilde{h}_{nm}(t) = Y_n^m[\boldsymbol{\theta}(t)]\, h(t), \tag{6}$$

where $Y_n^m(\boldsymbol{\theta})$ are the N3D-normalized, real-valued spherical harmonics of order n and degree m, cf. [23], evaluated at the direction $\boldsymbol{\theta}$, and the maximum order $n \leq N$ can be chosen freely. In the late diffuse part of the response, the implicit assumption of there being only a single DOA per time sample does not hold. As a result, a fluctuation of the DOA $\boldsymbol{\theta}(t)$ causes amplitude modulation and destroys narrow-band spectral content in $\tilde{h}_{nm}(t)$; typically, the longer low-frequency reverberation tails are hereby mixed towards higher frequencies, causing unnaturally long reverberation there [12,36] at high orders, cf. solid lines in Figure 5. However, theoretically, the expected temporal energy decay in an ideal (isotropic) diffuse field must be identical for any receiver of random-energy-efficiency-normalized directivity, such as the spherical harmonics, also after decomposition into frequency bands, and hence requires correction.

Despite the mismatch, formal derivation in [22] showed that quadratic summation across same-order spherical harmonics is omnidirectional $\sum_m |Y_n^m(\boldsymbol{\theta})|^2 = \frac{2n+1}{4\pi}$. Hereby, ASDM-upscaled ARIRs at least displays consistent broadband energies $\sum_m |\tilde{h}_{nm}(t)|^2 = \frac{2n+1}{4\pi}|h(t)|^2$ across all spherical harmonic orders n, for any sound field. To enforce consistency with spectral squares of $h(t)$, third-octave filtering is useful, where the bth sub-band signal $F_b\{h(t)\}$ with center frequency f_b is obtained from a bank of zero-phase filters F_b that is perfectly reconstructing $h(t) = \sum_b F_b\{h(t)\}$.

Appl. Sci. **2020**, *10*, 1631

For every sub band b and order n, an energy decay of the ASDM-upscaled ARIR $F_b\{\tilde{h}_{nm}(t)\}$ matching with the original one of $F_b\{h(t)\}$ is enforced by envelope correction

$$F_b\{h_{nm}(t)\} = F_b\{\tilde{h}_{nm}(t)\}\; w_n^b(t), \tag{7}$$
$$\text{with } w_n^b(t) = \sqrt{\tfrac{2n+1}{4\pi}} \sqrt{\frac{F_T\{F_b\{h(t)\}^2\}}{\sum_m F_T\{F_b\{h_{nm}(t)\}^2\}}},$$

where $F_T\{\cdot\}$ denotes temporal averaging with a time constant T (e.g., 100 ms). The energy decay reliefs for the initial and corrected result of ASDM are exemplary shown for a third octave $f_b = 2$ kHz and within the orders $n = \{1, 3, 5, 7\}$ in Figure 5.

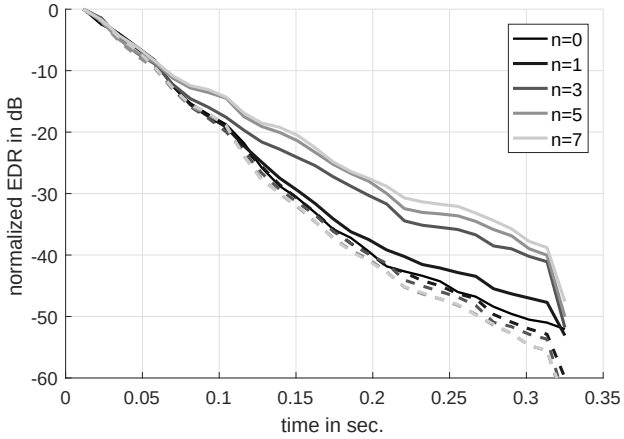


Figure 5. Energy decay relief (EDR) in a third-octave band with center frequency of 2 kHz. Solid and dashed lines indicate the order partitioned EDR before and after equalization as defined in Equations (6) and (7), respectively.

Finally, the ear signals are obtained by a convolution of the rotated Ambisonic signals with any state-of-the-art FIR binaural Ambisonic renderer, cf. Figure 4. Our study employs the time-invariant filters of the MagLS (magnitude-least-squares) renderer (The MagLS renderer is part of the IEM plugin suite which can be found here https://plugins.iem.at/) defined in [24,37] to get high-quality Ambisonic rendering already with an order $N = 3$. The filters were designed using a magnitude-learst-squares optimization that disregards phase match in favor of an improved HRTF magnitude at high frequencies, and hereby avoids spectral artifacts. MagLS also includes an interaural covariance correction that offers an optimal compromise to render diffuse fields consistently [23].

3.2. Listening Experiment: Design

Similar to *Experiment I*, listeners were asked to rate the spatial mapping, coloration, and continuity compared to the 1° reference. The test conditions included ARIR rendering with the ASDM target orders $N = \{1, 3, 5\}$ as well as the corresponding MOBRIR resolutions $\Delta\varphi = \{60, 45, 30\}^\circ$ rendered with LISHPh and broadband linear interpolation. Note that the set of MOBRIR resolutions is derived from the number of loudspeakers used for Ambisonics reproduction [23] in practice $\Delta\varphi_N \approx \frac{180^\circ}{N+1}$; for $N = 1$, we chose 60° instead of 90° to maintain a reasonable MOBRIR resolution.

We asked to rate the *timbre differences*, and *consistency of spatial mapping* for the five static listener orientations $\varphi = \{0, -35, 12, -15, 22.5\}^\circ$ which were switched automatically in 900 ms intervals and are restarted at the beginning of every audio loop. The orientations where chosen such that $0.25 < \alpha < 0.83$ for all resolutions in order to test for high interpolation depth; $\varphi = 0^\circ$ is included as reference orientation, and it marks the start of each loop. As source positions, we used a frontal and lateral virtual loudspeaker, cf. loudspeakers 3 and 7 in Figure 1a. The *continuity* test to compare dynamic rendering was similar as in *Experiment I*; see Section 2.2.

Appl. Sci. **2020**, *10*, 1631

3.3. Listening Experiment: Implementation

While ARIR-based rendering could be implemented in a multichannel DAW (e.g., Reaper) by using freely available convolution (http://www.matthiaskronlachner.com/?p=1910), rotation, and rendering plug-ins (https://plugins.iem.at/), there is no tested and easy-to-use plug-in for MOBRIR rendering, yet. To rule out any effects due to different implementations, we used the *pure data* implementation as described in Section 2.3 to also emulate ARIR-based rendering. To this end, we evaluated the ARIR BRIRs according to [22] to get a $\Delta\varphi = 1^\circ$ MOBRIR (for each ear)

$$AMOBRIR_q(t) = \sum_{\tau=0}^{T-1} \sum_{n=0}^{N} \sum_{m=-n}^{n} b_{nm}(\tau) \sum_{m'=-n}^{n} r_n^{mm'}(q\Delta\varphi)\, h_{nm'}(t-\tau), \tag{8}$$

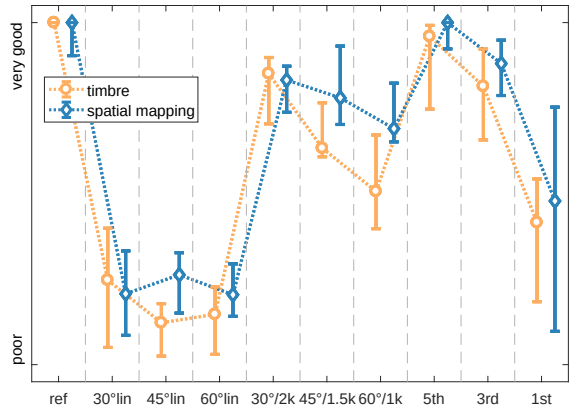
where $b_{nm}(t)$ is the FIR binaural Ambisonic renderer, $h_{nm'}(t)$ is the ARIR of the order N, q is the orientation index, and $r_n^{mm'}$ is the Ambisonic rotator. As binaural renderer $b_{nm}(t)$, we employed the one from [37] with KU100 HRIRs measured for 2702 directions [21]. The resulting AMOBRIR was linearly interpolated like the reference condition.

For playback, we again used AKG K702 headphones equipped with the IEM headtracker [8] and the experiment was conducted in a quiet office room.

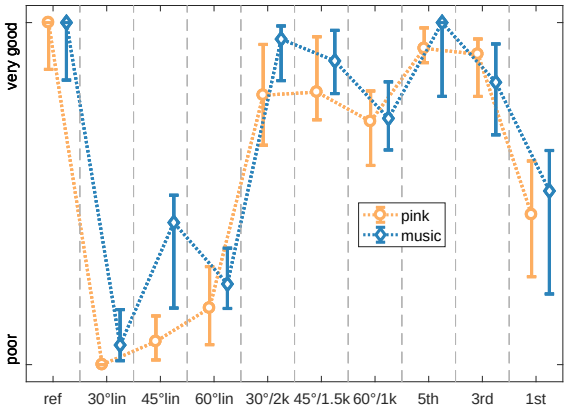
3.4. Results and Discussion

The results of the listening experiments with nine participants (all male experienced listeners with normal hearing, and between an age of 27 and 57) are depicted in Figure 6 and are discussed in detail below.

Timbre: While most of the conditions are rated significantly poorer than the reference, the following are not: LISHPh with 30°/2k and ARIR rendering with the ASDM-upscaled orders 3 and 5, cf. upper triangle in Table 3. ARIR rendering with 5th order is generally rated best; however, it is not significantly different to the 30°/2k and 3rd-order conditions. The timbral quality decreases with both orientation resolution and lower order, and thus the 60°/1k and 1st-order conditions yield significantly lower quality. The broadband linear interpolation conditions received the poorest ratings and significantly differ from all other conditions, with the exception of 60°/1k and 1st order.



(a) Timbre and spatial mapping of the pooled data for all tested directions for a frontal source and head orientations of $\varphi = \{0, -35, 12, -15, 22.5\}^\circ$.



(b) Continuity of the pooled data (virtual speaker at front and side, respectively).

Figure 6. Median (markers) and 95% confidence interval (solid lines) of ratings from all nine subjects for testing the perceived difference to the reference (linearly interpolated BRIRs on a 1° resolution). Settings of the algorithm are indicated by $\Delta\varphi/f_c$, where *lin* denotes a broadband linear interpolation.

Appl. Sci. **2020**, *10*, 1631

Spatial Mapping: While the general trend is similar to the *timbre* results, for *spatial mapping* only the 1st order and all linear interpolations are significantly poorer than the reference, cf. lower triangle in Table 3. Again, the 5th-order ARIR rendering is rated highest, albeit not significantly better than the 3rd-order ARIR and all LISHPh conditions ($p > 0.68$). The 30°, 45°, and 60° linear interpolations are significantly outperformed by all other conditions except the 1st-order ARIR rendering. Please note that we tested static directions different from *Experiment I* and even though the trend is similar to Figure 3(a), the 30°/2k is not significantly different from the reference condition here. This can be addressed to participants not always rating the reference highest in *Experiment II*.

Table 3. *p*-values (Wilcoxon signed-rank test with Bonferroni–Holm correction) for ratings of timbre and spatial mapping of Experiment II. The upper triangle corresponds to timbre, the lower triangle to spatial mapping. Insignificant differences (*p*-values ≥ 0.05) are indicated by bold numbers.

Method	ref	**30°lin**	**45°lin**	**60°lin**	**30°/2k**	**45°/1.5k**	**60°/1k**	**5th**	**3rd**	**1st**
ref	-	0.01	0.01	0.01	**0.10**	0.01	0.01	**0.22**	**0.18**	0.01
30°lin	0.01	-	**0.36**	**1.55**	0.01	**0.06**	**0.25**	0.01	0.03	**1.14**
45°lin	0.01	**2.46**	-	**1.55**	0.01	0.01	0.01	0.01	0 .01	**0.13**
60°lin	0.01	**2.53**	**2.83**	-	0.01	0.01	0.01	0.01	0.01	**0.60**
30°/2k	**1.02**	0.01	0.01	0.01	-	**0.22**	0.03	**1.30**	**0.95**	**0.07**
45°/1.5k	**1.06**	0.01	0.01	0.01	**1.76**	-	**0.07**	**0.07**	**0.76**	0.04
60°/1k	**0.71**	0.00	0.00	0.00	0.04	**0.97**	-	0.03	**0.07**	**0.50**
5th	**2.54**	0.01	0.01	0.01	**1.05**	**0.97**	**0.68**	-	**0.56**	0.02
3rd	**0.97**	0.01	0.01	0.01	**2.66**	**2.83**	**1.06**	**1.05**	-	0.01
1st	0.01	**1.11**	**1.10**	**0.68**	0.03	**0.19**	**0.55**	0.01	0.03	-

Continuity: The ratings of the *continuity*, i.e., robustness of source position and timbre to head rotations, are depicted in Figure 6b and Table 4, respectively. Tendentially, quality ratings are higher for music compared to pink noise as source signal. Independent of the source signal, the 5th-order and 3rd-order ARIR conditions as well as all LISHPh conditions do not significantly differ from the reference condition ($p > 0.15$). Again, all linearly interpolated conditions and the 1st-order condition perform poorly, are significantly different to all other conditions, and are similar to each other.

Table 4. *p*-values (Wilcoxon signed-rank test with Bonferroni–Holm correction) for ratings of continuity in Experiment II. The upper triangle corresponds to the pink noise, the lower triangle to music as source signal. Insignificant differences (*p*-values ≥ 0.05) are indicated by bold numbers.

method	ref	**30°lin**	**45°lin**	**60°lin**	**30°/2k**	**45°/1.5k**	**60°/1k**	**5th**	**3rd**	**1st**
ref	-	0.01	0.01	0.01	**0.39**	**1.72**	**0.15**	**1.43**	**1.99**	0.03
30°lin	0.01	-	0.00	0.03	0.00	0.01	0.01	0.01	0.01	0.01
45°lin	0.01	0.01	-	**0.15**	0.01	0.01	0.01	0.01	0 .01	**0.06**
60°lin	0.01	**0.19**	**0.64**	-	0.01	0.01	0.01	0.01	0.01	**0.08**
30°/2k	**1.67**	0.01	0.01	0.01	-	**1.57**	**1.89**	**0.98**	**1.99**	**0.05**
45°/1.5k	**1.86**	0.01	0.01	0.01	**0.39**	-	**1.44**	**1.72**	**1.89**	0.02
60°/1k	**0.08**	0.01	**0.05**	0.01	**0.09**	**0.26**	-	**0.16**	**0.63**	**0.16**
5th	**1.86**	0.01	0.01	0.01	**1.17**	**1.72**	**0.12**	-	**1.61**	0.01
3rd	**0.08**	0.01	0.04	0.01	**1.34**	**1.67**	**1.66**	**0.15**	-	0.03
1st	0.01	**0.12**	**1.72**	**0.37**	0.01	0.02	0.02	0.01	**0.07**	-

Appl. Sci. **2020**, *10*, 1631

4. Conclusions

We evaluated two fundamentally different measurement-based binaural audio rendering strategies in a novel comparative listening experiment: The dummy-head-based strategy employs binaural impulse responses measured in multiple orientations (MOBRIRs) and hereby contains the required set of binaural cues of the dummy head for dynamic (head-tracked) rendering. The Ambisonics-based strategy uses the room impulse response measured by a first-order Ambisonic microphone (ARIR) in a single orientation, which is upscaled from its weak directional resolution to higher orders using the Ambisonic spatial decomposition method (ASDM). Dynamic binaural rendering is then accomplished separately through an Ambisonic rotator and binaural renderer.

Our experiment successfully compared the perceptual performance of both strategies, for static rendering in terms of timbre and spatial mapping, and for dynamic rendering concerning the resulting temporal continuity, overall. We found that the 5th-order Ambisonics-based rendering strategy (ASDM) outperformed the dummy-head-based rendering for resolutions coarser than 30°. By this and by its clear separation of room-related from head-related aspects, we consider ASDM binaural rendering as the versatile high-quality option for dynamic binaural rendering based on measured room responses. We published audio examples, an example implementation, and all experimental response data for reproducible research.

Concerning the dummy-head-based strategy with MOBRIRs, we summarized the analysis and made available all experimental response data from previous experiments [20]. The results indicate that linear interpolation between the different dummy-head orientations is always outperformed by the linear interpolation and switched high-frequency phase method (LISHPh). This processing strategy achieved a convincing rendering quality with an orientation resolution of 15° and 30°, when compared to a 1° linearly interpolated reference.

The underlying RIR measurements were taken from the IEM production studio with a reverberation time of $T_{60} \approx 0.4\,s$. This specific choice of room was found suitable to study the MOBRIR interpolation and ASDM binaural rendering as its pronounced direct and early parts are expected to be critical considering specific timbral or spatial mapping deficiencies, and its reverberation is suitable to evoke externalized impressions in typical office environments. Note that neither of the investigated rendering strategies was specifically optimized for the specific room and signals. Although not formally tested, we assume the results to hold for a variety of other acoustic environments. An example patch and a set of more reverberant BRIRs (RT=2.8s) are provided online (https://phaidra.kug.ac.at/o:100863).

Acknowledgments: The authors thank all listeners for their participation in the listening experiment.

Author Contributions: Conceptualization, M.Z. and F.Z.; Writing-Original Draft Preparation M.Z. and F.Z. with periodic contributions by M.F.; Writing—Review & Editing, M.Z. and F.Z.; Software, M.Z. with periodic contributions by F.Z.; Listening Experiment Implementation, F.Z. and M.Z. with periodic contributions by M.F.; Listening Experiment Design, M.Z., M.F., and F.Z.; Data Analysis, M.F. and M.Z.; Measurements, M.Z. and F.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appl. Sci. **2020**, *10*, 1631

Appendix A. MOBRIR PureData Real-Time Processing Patch

An example patch for the pure-data real-time processing environment using $M = 7$ orientations with a resolution of $\Delta\phi = 15°$ can be accessed here https://phaidra.kug.ac.at/o:97087. The patch for the high-frequency processing is exemplary shown for seven orientations in Figure A1.

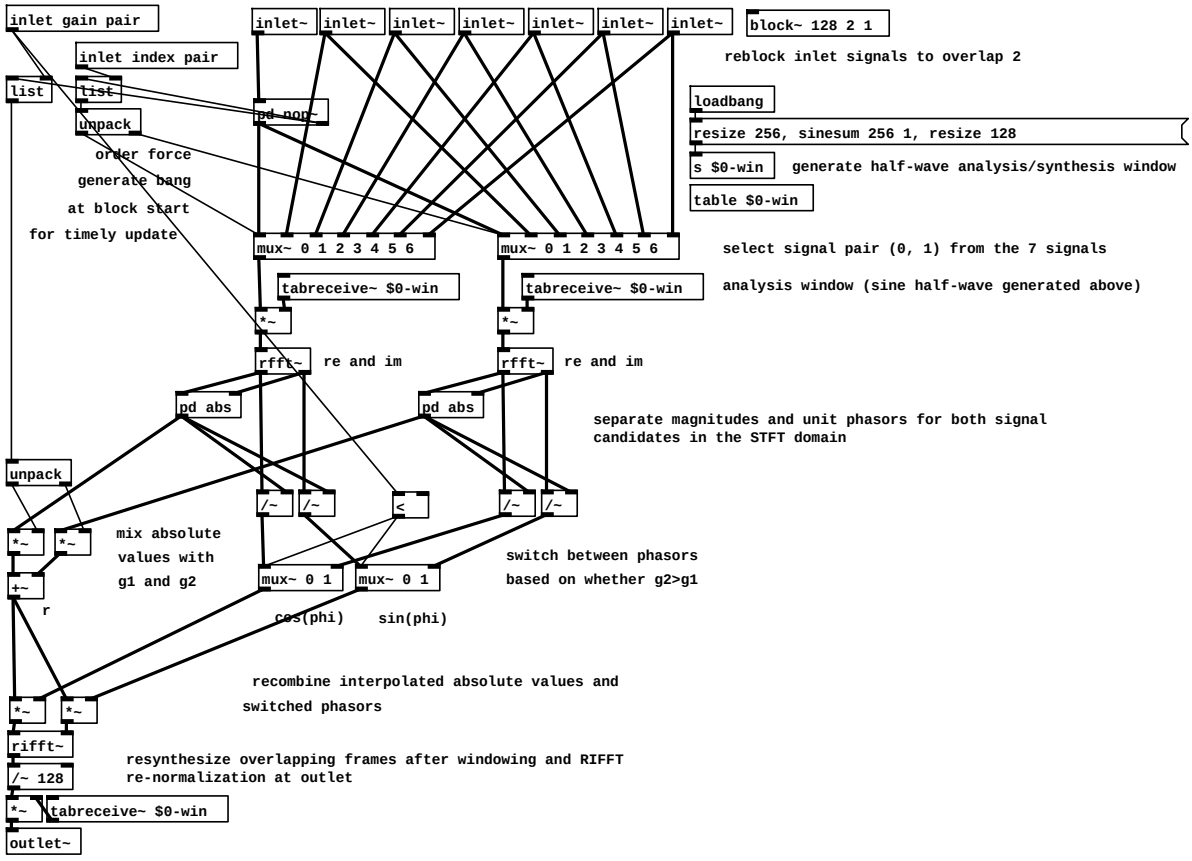


Figure A1. Implementation of the high-frequency patch in Pd.

Appendix B. ASDM MATLAB Source Code

The MATLAB source code of the proposed ASDM method can be found in Listing 1. Please note that you need to install the spherical harmonic transform which can be accessed from https://github.com/polarch/Spherical-Harmonic-Transform.

```
%% First order B-format RIR (ARIR)
[x,fs] = audioread(fname); % sorted in W,X,Y,Z - fs is the sampling frequency

%% Parameters and Settings
N = 3; % Ambisonics order
Nfft = 2^ceil(log2(size(x,1))); % fft length
Lsmooth_dirfluct = 17;
win_dirfluct=hann(Lsmooth_dirfluct);
Lsmooth_specdecay = 4096; % smoothing for fs = 44.1kHz
win_specdecay=hann(Lsmooth_specdecay);

%% PIV DOA estimation
[b,a] = butter(4,[200 3000]/(fs/2)); % bandpass with fl = 100Hz and fh = 3kHz
xbp = filtfilt(b,a,x);
e  = xbp(:,1).^2;          e = circshift(fftfilt(win_dirfluct, e),-floor(Lsmooth_dirfluct/2));
ix = xbp(:,1).*xbp(:,2);   ix = circshift(fftfilt(win_dirfluct,ix),-floor(Lsmooth_dirfluct/2));
```

```
iy = xbp(:,1).*xbp(:,3);   iy = circshift(fftfilt(win_dirfluct,iy),-floor(Lsmooth_dirfluct/2));
iz = xbp(:,1).*xbp(:,4);   iz = circshift(fftfilt(win_dirfluct,iz),-floor(Lsmooth_dirfluct/2));
azi = atan2(iy, ix);
zen = atan2(sqrt(ix.^2+iy.^2),iz);

%% ASDM Upscaling of the ARIR
Ysh = getSH(N,[azi,zen],'real'); % from https://github.com/polarch/Spherical-Harmonic-Transform/
x = x(:,1).*Ysh; % upmixing

%% Spectral Decay Correction
H = thirdoctave_filter_bank_linph(Nfft,fs);
x_c = zeros(Nfft,(N+1)^2); % corrected upscaled ARIR
x_c(1:size(x,1),1) = x(:,1);
for k = 1:size(H,2)
    xthird0 = ifft(fft(x(:,1),Nfft).*H(:,k));
    xthird0rms = sqrt(circshift(fftfilt(win_specdecay,xthird0.^2),-Lsmooth_specdecay/2));
    for n = 1:N
        nidx = n^2+(1:2*n+1);
        xthirdn = ifft(fft(x(:,nidx),Nfft).*H(:,k));
        xthirdnrms = sqrt(sum(circshift(fftfilt(win_specdecay,xthirdn.^2),-Lsmooth_specdecay/2),2));
        w_c = xthird0rms./(xthirdnrms+1e-6)*(2*n+1); % correction window
        x_c(:,nidx) = x_c(:,nidx)+xthirdn.*w_c;
    end
end
[n,m] = shindex(N);
renormalize = 1./sqrt(2*n+1);
x_c = x_c(1:size(x,1),:) * diag(renormalize);

%% Function Definitions
function [n,m] = shindex(nmax)
k = 0:(nmax+1)^2-1;
n = floor(sqrt(k));
m = k-n.^2-n;
end

function H = thirdoctave_filter_bank_linph(Nfft,fs)
f=linspace(0,fs/2,Nfft/2+1);
f(1)=f(2)/4;
fc = 25*2.^(0:1/3:9.9); %third-octave vector
H = zeros(Nfft/2+1,length(fc));
for k = 1:length(fc)
    nthoctaves = log2(f/fc(k))*3; % 3rd-octaves distance from center freq.
    upper = 1.0*(k<length(fc)); % upper 3rd-octave limit (high-pass in last band)
    lower = -1.0*(k>1); % lower 3rd-octave limit (low-pass in first band)
    nthoctaves = max(min(nthoctaves,upper,lower);
    H(:,k) = cos(nthoctave*pi/2).^2;
end
H = [H;flipud(H(2:end-1,:))];
end
```

Listing 1: MATLAB source code of the proposed ASDM method.

References

1. Møller, H. Fundamentals of binaural technology. *Appl. Acoust.* **1992**, *36*, 171–218, doi:10.1016/0003-682X(92)90046-U.
2. Pollack, I.; Trittipoe, W. Binaural listening and interaural noise cross correlation. *J. Acoust. Soc. Am.* **1959**, *31*, 1250–1252, doi:10.1121/1.1907852.
3. Okano, T.; Beranek, L.L.; Hidaka, T. Relations among interaural cross-correlation coefficient (IACCE), lateral fraction (LFE), and apparent source width (ASW) in concert halls. *J. Acoust. Soc. Am.* **1998**, *104*, 255–265, doi:10.1121/1.423955.

4. Lindau, A. Binaural Resynthesis of Acoustical Environments-Technology and Perceptual Evaluation. Ph.D. Thesis, TU Berlin, 2014.
5. Smyth, S.M. Personalized headphone virtualization. U.S. Patent 7,936,887, 3 May 2011.
6. Smyth, S.; Smyth, M.; Cheung, S. Headphone Surround Monitoring for Studios. In Proceedings of the Convention of the Audio Engineering Society, London, UK, 9 April 2008; pp. 1–7.
7. Satongar, D.; Lam, Y.W.; Pike, C. Measurement and Analysis of a Spatially Sampled Binaural Room Impulse Response Dataset. In Proceedings of the 21st International Congress on Sound and Vibration, Beijing, China, 13–17 July 2014; pp. 1–8.
8. Romanov, M.; Berghold, P.; Rudrich, D.; Zaunschirm, M.; Frank, M.; Zotter, F. Implementation and evaluation of a low-cost head-tracker for binaural synthesis. In Proceedings of the Convention of the Audio Engineering Society, Berlin, May 2017.
9. Begault, D.R.; Wenzel, E.M.; Anderson, M.R. Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source. *J. Audio Eng. Soc.* **2001**, *49*, 904–916.
10. Brungart, D.S.; Kordik, A.J.; Simpson, B.D. Effects of headtracker latency in virtual audio displays. *J. Audio Eng. Soc.* **2006**, *54*, 32–44.
11. Hendrickx, E.; Stitt, P.; Messonnier, J.C.; Lyzwa, J.M.; Katz, B.F.; de Boishéraud, C. Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis. *J. Acoust. Soc. Am.* **2017**, *141*, 2011–2023, doi:10.1121/1.4978612.
12. Zaunschirm, M.; Baumgartner, C.; Schörkhuber, C.; Frank, M.; Zotter, F. An Efficient Source-and-Receiver-Directional RIR Measurement Method. *Fortschritte der Akustik - DAGA* **2017**, pp. 1343–1346.
13. Pörschmann, C.; Wiefling, S. Perceptual Aspects of Dynamic Binaural Synthesis based on Measured Omnidirectional Room Impulse Responses. In Proceedings of the International Conference on Spatial Audio, Seattle, WA, USA, 26–30 May 2015.
14. Menzer, F. Binaural Audio Signal Processing Using Interaural Coherence Matching. Ph.D. Thesis, EPFL, Lausanne, Switzerland, 2010.
15. Lindau, A.; Maempel, H.J.; Weinzierl, S. Minimum BRIR grid resolution for dynamic binaural synthesis. *J. Acoust. Soc. Am.* **2008**, *123*, 3498, doi:10.1121/1.2934364.
16. Algazi, V.R.; Duda, R.O.; Thompson, D.M. Motion-tracked binaural sound. *J. Audio Eng. Soc.* **2004**, *52*, 1142–1156.
17. Pruša, Z.; Balazs, P.; Søndergaard, P.L. A Non-iterative Method for STFT Phase (Re)Construction. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **2016**, *25*, 1154–1164.
18. Hom, R.C.M.; Algazi, V.R.; Duda, R.O. High-Frequency Interpolation for Motion-Tracked Binaural Sound. In Proceedings of the Convention of the Audio Eng. Soc. 121, San Francisco, October 2006.
19. Lindau, A.; Roos, S. Perceptual evaluation of discretization and interpolation for motion-tracked binaural (MTB) recordings. In Proceedings of the 26th Tonmeistertagung, Leipzig, Germany, 25–28 November 2010; pp. 680–701.
20. Zaunschirm, M.; Frank, M.; Franz, Z. Perceptual Evaluation of Variable-Orientation Binaural Room Impulse Response Rendering. In Proceedings of the Conference of the Audio Eng. Soc.: 2019 AES International Conference on Immersive and Interactive Audio, York, March 2019.
21. Bernschütz, B. A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100. *Fortschritte der Akustik - AIA-DAGA*, Merano, Italy, 18–21 March 2013, 592–595.
22. Zaunschirm, M.; Frank, M.; Zotter, F. BRIR Synthesis Using First-Order Microphone Arrays. In Proceedings of the Conference of the Audio Eng. Soc. 144,Milan,May 2018; pp. 1–10.
23. Zotter, F.; Frank, M. *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*; SpringerOpen: Berlin, Germany, 2019; doi:10.1007/978-3-030-17207-7.
24. Zaunschirm, M.; Schörkhuber, C.; Höldrich, R. Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint. *J. Acoust. Soc. Am.* **2018**, *143*, 3616–3627, doi:10.1121/1.5040489.
25. International Telecommunication Union. ITU-R BS.1534-3, Method for the subjective assessment of intermediate quality level of audio systems. *ITU-R Recommendation* **2015**, 1534.

26. Karjalainen, M.; Piirilä, E.; Järvinen, A.; Huopaniemi, J. Comparisonof Loudspeaker Equalization Methods Based on DSP Techniques. *J. Audio Eng. Soc.* **1999**, *47*, 14 – 31.
27. Lipshitz, S.P.; Vanderkooy, J. in-Phase Crossover Network Design. *J. Audio Eng. Soc.* **1986**, *34*, 889–894.
28. D'Appolito, J. Active realization of multiway all-pass crossover systems. *J. Audio Eng. Soc.* **1987**, *35*, 239–245.
29. Wilcoxon, F. Individual comparisons by ranking methods. In *Breakthroughs in Statistics*; Springer: Berlin, Germany, 1992; pp. 196–202.
30. Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian J. Stat.* **1979**, *6*, 65–70.
31. Rayleigh, L. On our perception of sound direction. *Philos. Mag. Ser. 6* **1907**, *13*, 214–232, doi:10.1080/14786440709463595.
32. Ivanic, J.; Ruedenberg, K. Rotation matrices for real spherical harmonies. direct determination by recursion. *J. Phys. Chem.* **1996**, *100*, 6342–6347, doi:10.1021/jp9833350.
33. Pinchon, D.; Hoggan, P.E. Rotation matrices for real spherical harmonics: General rotations of atomic orbitals in space-fixed axes. *J. Phys. A Math. Theor.* **2007**, *40*, 1597–1610, doi:10.1088/1751-8113/40/7/011.
34. Tervo, S.; Pätynen, J.; Kuusinen, A.; Lokki, T. Spatial decomposition method for room impulse responses. *J. Audio Eng. Soc.* **2013**, *61*, 17–28.
35. Jarrett, D.P.; Habets, E.A.P.; Naylor, P.A. 3D Source localization in the spherical harmonic domain using a pseudointensity vector. In Proceedings of the European Signal Processing Conference, Aalborg, Denmark, 23–27 August 2010; pp. 442–446.
36. Frank, M.; Zotter, F. Spatial impression and directional resolution in the reproduction of reverberation. *Fortschritte der Akustik - DAGA* Aachen, Germany, 14–17 March 2016, pp. 1304–1307.
37. Schörkhuber, C.; Zaunschirm, M.; Höldrich, R. Binaural Rendering of Ambisonic Signals via Magnitude Least Squares. *Fortschritte der Akustik - DAGA* Munich, Germany, 19–22 March 2018, *44*, 339–342.

6 Concluding Remarks

This PhD project dealt with the concept, and the hardware-efficient implementation for measurement-based auralization of rooms that allows to incorporate adjustable source and receiver directivities. The proposed source-and-receiver-directional Ambisonic room impulse response (SRD ARIR) capturing and processing method yields a resolution enhancement of low-order directional RIR measurements and its most hardware-efficient implementation employs a small set of first-order (minimum of 4 channels) measurements. The SRD ARIR method was evaluated in a comparative listening experiment to a variety of other auralization techniques which rely on more extensive measurements.

In the comparative listening experiment the IKO (icosahedral loudspeaker array), with 20 individual loudspeakers mounted on the surfaces of an icosahedron, was employed as source of high-order directivity. In the scope of this PhD project, a contribution to both the spherical harmonic beamforming capabilities and formalism as well as the documentation of the practical achievable beam patterns of the IKO have been made. While the IKO allows for controlling beams and synthesizing source directivities up to third order, its own directivity is described by orders up to 16 (evaluation of pressure measurements done with a surrounding microphone array).

An alternative to the proposed modular and interactive SRD ARIR method is a technique that relies on the measured multiple-orientation binaural room impulse response (MOBRIRs). The studies presented in this PhD project dealt with finding the optimal MOBRIR grid resolution as a trade-off between measurement-effort and obtained perceptual quality when rendering dynamically (i.e. accounting for the head movements of a listener). It was found that a MOBRIR resolution of 15° combined with the linear-interpolation and switched high-frequency phase method (LISHPh) yields rendering qualities comparable to a 1° linearly interpolated reference. Further, the optimal parameter settings (cross-over frequency, block size, etc.) of the the LISHPh method were discussed and a demo implementation as well as all underlying measurements and listening examples are available online[4].

Besides the MOBRIR-based auralization, all other tested auralization techniques are based on either a direct measurement of the Ambisonic RIRs (ARIRs) or a processing that yields the ARIRs. Either way, all the ARIR-based auralization techniques facilitate a binaural Ambisonics renderer. In the scope of this thesis rendering methods based on a frequency-dependent time alignment of head-related transfer functions (HRTFs) in pre-processing or a magnitude-least-squares optimization that disregards a phase-match in favor of a magnitude match at high frequencies were proposed. Both renderers optionally include an interaural covariance correction that offers to render diffuse fields consistently. The proposed renderers yield a representation of ear directivities with radically improved mapping of timbre at high frequencies. The findings from perceptual evaluations indicate that already an Ambisonics order of three allows for high-quality rendering.

[4] `https://phaidra.kug.ac.at/o:76741`

Finally, the results from the comparative study showed that the proposed SRD ARIR method performed similarly to the reference condition (MOBRIR-based) as no significant differences for the attributes localizability, direction, or distance have been found. With its benefits in terms of modularity and hardware efficiency, the SRD ARIR method is convenient for any application requiring a flexible exchange of source and receiver directivities while keeping the number of responses needed for the room description to a minimum.

Note that a collection of RIRs, listening experiment response data, the statistical analysis, high resolution array data, and example implementations are made available online[5].

[5] `https://phaidra.kug.ac.at/o:104417`