OPEN ACCESS Freely available online
J. M. Arend, S. V. A. Garí, C. Schissler, F. Klein, and P. W. Robinson, "Six-Degrees-of-Freedom Parametric Spatial Audio Based on One Monaural Room Impulse Response"
*J. Audio Eng. Soc.*, vol. 69, no. 7/8, pp. 557–575, (2021 July/August).
DOI: https://doi.org/10.17743/jaes.2021.0009

# Six-Degrees-of-Freedom Parametric Spatial Audio Based on One Monaural Room Impulse Response

**JOHANNES M. AREND,**[1 2]* *AES Student Member*,
(johannes.arend@th-koeln.de)

**SEBASTIÀ V. AMENGUAL GARÍ,**[3] *AES Associate Member*, **CARL SCHISSLER,**[3]
(samengual@fb.com) (cschissler@fb.com)

**FLORIAN KLEIN,**[4]* **AND PHILIP W. ROBINSON,**[3] *AES Member*
(florian.klein@tu-ilmenau.de) (philrob22@fb.com)

[1] *Institute of Communications Engineering, TH Köln - University of Applied Sciences, Cologne, D-50679, Germany*
[2] *Audio Communication Group, Technical University of Berlin, Berlin, D-10587, Germany*
[3] *Facebook Reality Labs Research, Redmond, WA 98052, USA*
[4] *Electronic Media Technology Lab, Technical University of Ilmenau, Ilmenau, D-98693, Germany*

Parametric spatial audio rendering is a popular approach for low computing capacity applications, such as augmented reality systems. However most methods rely on spatial room impulse responses (SRIR) for sound field rendering with 3 degrees of freedom (DoF), i.e., for arbitrary head orientations of the listener, and often require multiple SRIRs for 6-DoF rendering, i.e., when additionally considering listener translations. This paper presents a method for parametric spatial audio rendering with 6 DoF based on one monaural room impulse response (RIR). The scalable and perceptually motivated encoding results in a parametric description of the spatial sound field for any listener's head orientation or position in space. These parameters form the basis for the binaural room impulse responses (BRIR) synthesis algorithm presented in this paper. The physical evaluation revealed good performance, with differences to reference measurements at most tested positions in a room below the just-noticeable differences of various acoustic parameters. The paper further describes the implementation of a 6-DoF real-time virtual acoustic environment (VAE) using the synthesized BRIRs. A pilot study assessing the plausibility of the 6-DoF VAE showed that the system can provide a plausible binaural reproduction, but it also revealed challenges of 6-DoF rendering requiring further research.

## 0 INTRODUCTION

Augmented reality (AR) applications must provide perceptually plausible spatial audio rendering consistent with the real acoustic environment in order to create a coherent soundscape of real and virtual sources. In parallel, the rendering must be computationally lightweight, since the limited resources usually have to be shared with other computationally demanding components, such as visuals, sensors, and mapping, among others. One approach to meet these conflicting demands is parametric rendering. The sound field is first encoded offline into a parametric description covering the perceptually essential components and then decoded (in real-time) for efficient spatial audio reproduction scalable to the technical conditions. Furthermore parametric rendering offers a high degree of flexibility, as the parameters can be easily adjusted, for example, to represent different source-receiver conditions or room acoustic situations.

Various methods for the parametrization (encoding) and rendering (decoding) of sound fields based on spatial room impulse responses (SRIRs) have been presented (e.g., [1–6]). The methods usually split the sound field into directional and diffuse components, which are then processed and rendered separately. To determine the directional components, the approaches exploit the directional information of the SRIR to estimate the direction of arrival (DOA) of the direct sound and the early reflections.

The methods mentioned above usually decode the sound field only at the measurement point for the listener's head orientations, i.e., only for 3 degrees of freedom (DoF). Recent work extended those methods to render the sound field for 6 DoF and thus for arbitrary head orientations and room

*The work was done as a research intern at Facebook Reality Labs Research in Redmond, WA, USA.

positions of the listener. The methods typically require at least one or even multiple SRIRs from different positions in the room and derive a description of the sound field at arbitrary positions by extrapolation based on one SRIR (e.g., [7, 8]) or by interpolation between the distributed SRIR measurements (e.g., [9, 10]). For a comprehensive overview of 6-DoF rendering methods please refer to [10].

In the context of AR, however, it may be necessary to obtain a spatial parametric description of the sound field for 6-DoF rendering based on one monaural room impulse response (RIR). For example blind system identification used to estimate room acoustic parameters in real-time results in a parametric description of the monaural RIR, which needs to be further processed and finally decoded for parametric spatial audio reproduction. Advances in blind system identification could in the near future enable RIR estimation with consumer devices. Furthermore a monaural RIR can be measured relatively easily even with consumer equipment (such as a smartphone), which, if appropriately encoded and decoded, would allow users and content creators of AR applications to quickly and easily obtain a spatial audio reproduction corresponding to the real space. Moreover an enormous number of monaural RIRs are available to the public, which can be used to build a database of parametrically described spaces suitable for real-time AR audio rendering.

Encoding a monaural RIR provides common parameters such as the amplitude and the time of arrival (TOA) of direct sound and early reflections as well as the frequency-dependent reverberation time or the direct-to-reverberation ratio (DRR). However, since a monaural RIR does not contain directional information, the DOAs of the direct sound and the early reflections must, for example, either be predefined, pseudo-randomly assigned, or estimated based on the room geometry.

Pörschmann et al. [11] presented a first approach for synthesizing binaural room impulse responses (BRIRs) for any desired head orientation based on one measured RIR. Their method aims to decompose the broadband RIR into a directional and diffuse part and process them separately (quite similar to spatial impulse response rendering (SIRR), where several band-passed parts of a SRIR are decomposed into directional and diffuse components [1]). The directional part, consisting of direct sound and (grouped) early reflections, is estimated by reflection detection and synthesized by convolving small chunks of the RIR containing directional information with head-related impulse responses (HRIRs). The algorithm assigns a predefined DOA (and optionally TOA) to the direct sound according to the measurement setup while it assigns pseudo-randomized DOAs to the early reflections due to missing spatial information. The binaural diffuse reverberation is synthesized by convolving small chunks of the RIR with chunks of binaural white noise and summing with overlap-add, which is essentially a de-correlation of the RIR.

This paper presents a novel approach for parametric spatial audio with 6 DoF based on one monaural RIR. The method, which we named *Paraspax* (PARAmetrization, SPAtialization, and eXtrapolation of monaural room impulse responses), is inspired by the method introduced by Pörschmann et al. but has significant extensions and improvements.

Through encoding of the monaural RIR, the present method derives monaural and spatial parameters suitable for parametric BRIR synthesis or real-time parametric rendering. The approach provides a scalable reflection detection, which allows selecting a specific number of perceptually salient reflections [12] for processing and position-dynamic rendering. The DOAs for the selected early reflections can be derived in different ways: a) based on a pseudo-randomized directional distribution; b) based on a simple image source model (ISM) for a shoebox-shaped room, which approximates the real room's geometry; or c) based on a previously determined DOA pattern obtained, for example, by applying the spatial decomposition method (SDM) [3] to SRIR measurements. Thus the latter option makes it possible to combine SRIR measurements with the scalable encoding of the presented approach.

Furthermore the proposed method allows filtering the direct sound according to the sound source directivity, thus enhancing the presentation when, for example, walking around a virtual sound source. Assuming a shoebox-shaped room with the selected early reflections as image sources, the parameters can be extrapolated to any position in the room, which allows 6-DoF parametric spatial audio rendering.

We further present a processing chain for parametric BRIR synthesis based on the (extrapolated) parameters and measured monaural RIR. The method decomposes the sound field into directional and diffuse components. In contrast to previously described approaches, which also perform a decomposition (e.g., [1, 2]), the *Paraspax* method does not require spatial information (i.e., no multichannel RIR) for the decomposition of the sound field but achieves it solely based on a monaural RIR. In the synthesis, BRIRs are constructed by recombining the directional and diffuse components, which are adjusted according to the parameters.

In the paper we evaluate the synthesized BRIRs by comparison with measurements. Moreover we present an implementation of a 6-DoF virtual acoustic environment (VAE) in a selected room, based on BRIRs synthesized with the *Paraspax* method and a purpose-built real-time framework, which can be used to perform psychoacoustic experiments for research on AR. Lastly we present results of a pilot study using the 6-DoF framework to examine the perceptual plausibility of the *Paraspax* method.

## 1 *PARASPAX* METHOD

*Paraspax* can be grouped into three basic processing blocks for encoding: parametrization, spatialization, and extrapolation, as shown in Fig. 1. The parametrization provides standard monaural room acoustic parameters, the amplitude and TOA of direct sound and early reflections using a reflection detection algorithm, the magnitude responses of reflection filters, and the reverberation level. The spa-

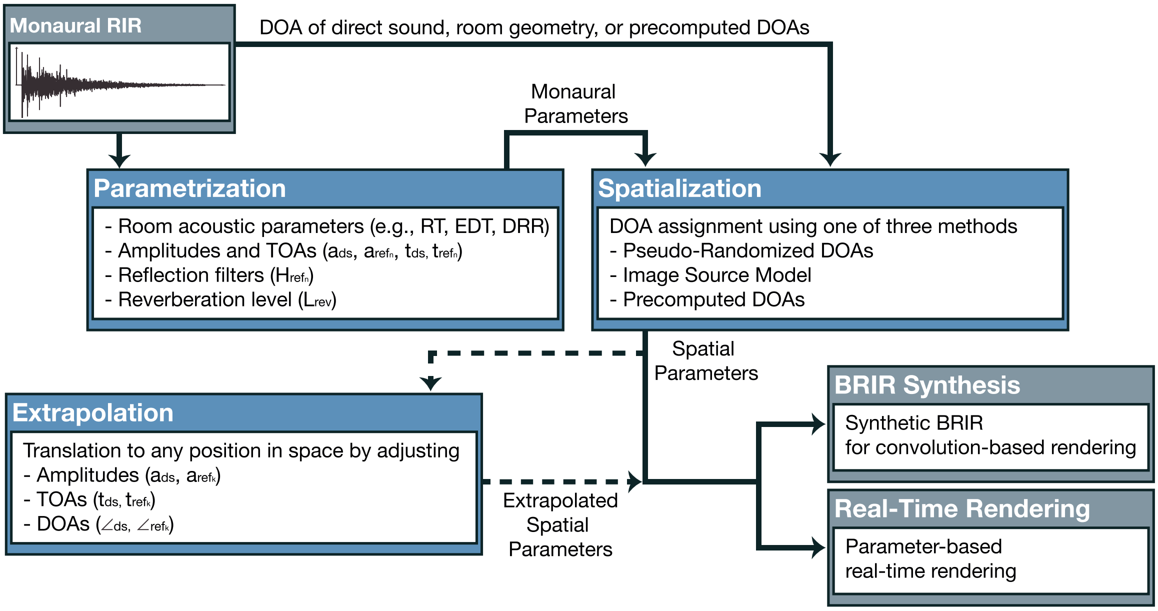
Monaural RIR
DOA of direct sound, room geometry, or precomputed DOAs
Monaural Parameters
Parametrization
- Room acoustic parameters (e.g., RT, EDT, DRR)
- Amplitudes and TOAs ($a_{ds}$, $a_{ref_n}$, $t_{ds}$, $t_{ref_n}$)
- Reflection filters ($H_{ref_n}$)
- Reverberation level ($L_{rev}$)
Spatialization
DOA assignment using one of three methods
- Pseudo-Randomized DOAs
- Image Source Model
- Precomputed DOAs
Spatial Parameters
Extrapolation
Translation to any position in space by adjusting
- Amplitudes ($a_{ds}$, $a_{ref_x}$)
- TOAs ($t_{ds}$, $t_{ref_x}$)
- DOAs ($\angle_{ds}$, $\angle_{ref_x}$)
Extrapolated Spatial Parameters
BRIR Synthesis
Synthetic BRIR for convolution-based rendering
Real-Time Rendering
Parameter-based real-time rendering

Fig. 1. Block diagram of the *Paraspax* method. The parametrization of the monaural RIR provides basic monaural room acoustic parameters, amplitudes and TOAs of the direct sound and early reflections, the magnitude responses of the reflection filters, and the reverberation level. The spatialization assigns DOAs to the direct sound and the selected reflections using one of the three implemented methods. The optional extrapolation allows for listener translations by adjusting amplitudes, TOAs, and DOAs of direct sound and early reflections. The parametric description can be used for BRIR synthesis or real-time parametric rendering.

tialization assigns DOAs to the direct sound and selected early reflections, based on a pseudo-randomized directional distribution, a simple ISM, or precomputed DOAs. Extrapolation allows listener translations to any position in space by adjusting amplitudes, DOAs, and TOAs of direct sound and early reflections. The parametric description can then either be used to synthesize BRIRs or passed to a real-time rendering engine. The following sections describe the encoding and BRIR synthesis in detail. A Matlab-based implementation of the proposed method with sample code is available online.[1]

## 1.1 Parametrization

Standard room acoustic parameters according to ISO 3382-2 are calculated in a first step based on the pressure response $p$, i.e., the monaural RIR, such as reverberation time ($\mathrm{RT}_{20/30/60}$), early decay time (EDT), clarity ($C_{50/80}$), directness ($D_{50/80}$), and energy decay curves (EDC) in octave or 1/3-octave frequency bands, as well as the broadband DRR and the mixing time. The latter is calculated according to the approach proposed by Abel et al. [13], which follows the assumption that the sound pressure amplitudes in a diffuse sound field assume a Gaussian distribution. For this the normalized echo density profile is calculated (window length of about 21 ms, as recommended by the authors), which describes to what extent the amplitude distribution of an RIR approximates a Gaussian distribution over time. As diffuse energy increases, the echo density profile of an RIR rises, and the mixing time is defined as when the echo density profile reaches the value of one the first time.

In further processing described in the following sections, the TOA and amplitude of direct sound ($t_{ds}$, $a_{ds}$) and the $n$ detected early reflections ($t_{ref_n}$, $a_{ref_n}$), as well as the magnitude response of the reflection filters for each detected reflection ($H_{ref_n}$) and the reverberation level ($L_{rev}$), are estimated.

### 1.1.1 *Direct Sound*

The TOA of the direct sound $t_{ds}$ is determined using onset-detection with a threshold of −20 dB in relation to the maximum value of the 10 times upsampled RIR. The amplitude of the direct sound $a_{ds}$ is calculated as the root-mean-squared (RMS) average of an asymmetrical window centered around $t_{ds}$, starting 0.5 ms before and ending 1 ms after $t_{ds}$, as proposed by Brinkmann et al. [12]. Starting slightly before $t_{ds}$ accounts for pre-ringing artifacts, and the length of 1 ms after $t_{ds}$ relates to the time frame in which summing localization takes place, i.e., the time in which multiple coherent sound sources are perceived as one auditory event [14, ch. 3.1].

### 1.1.2 *Early Reflections*

The reflection detection can be performed on the entire RIR or in a limited time range, e.g., up to twice the calculated mixing time, to estimate the TOAs of the early reflections only in the early part of the RIR. The selection process is motivated by perceptual mechanisms and is described below.

The RIR is windowed with a sliding rectangular window of 1 ms, and the TOA of a reflection is defined as the time index where the local energy is 3 times higher than the median energy in the window [15]. The window size

[1] Available: https://github.com/facebookresearch/Paraspax

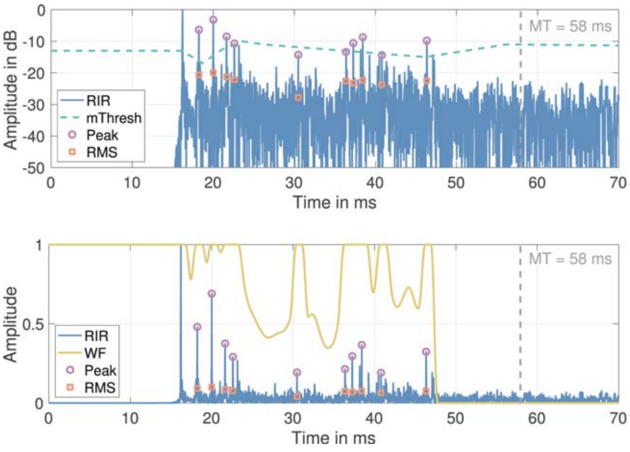

Fig. 2. Result of the reflection detection with $k = 10$ loudest reflections, peak and RMS amplitude values, adapted masking threshold (top), and estimated weighting function (bottom) for the RIR decomposition.

ensures a high temporal resolution in order to capture the perceptually important floor reflection [16, 17]. For each detected reflection, the RMS amplitude $a_{\mathrm{ref}_n}$ in an asymmetrical window of 1.5 ms (similar as described above for the direct sound), as well as the peak amplitude, is calculated.

The reflection selection method can lead to multiple detected reflections that are very close in time. To avoid this a next step evaluates whether another reflection occurs in a time range of 1 ms after a detected reflection (we choose this time range again according to mechanisms of summing localization). If this is the case, the reflection with the higher amplitude is declared valid, while the reflection with the lower amplitude is excluded from the selection. Optionally, in a further selection step, the peak amplitude of the detected reflections can be compared to the reflection masking threshold determined by Olive and Toole [18]. This procedure links the reflection detection even more closely to auditory perception and makes it possible to exclude potentially inaudible reflections from dynamic rendering.

Finally the selected reflections are sorted according to their amplitude and the $k$ loudest early reflections are selected for spatialization and dynamic reproduction. Although other sorting and selection mechanisms (e.g., by time order) are also conceivable, previous studies have shown that rendering the loudest $k$ (with $k = 6$–$10$) reflections is a valid method to reproduce the most perceptually important and salient reflections [19, 12].

Fig. 2 shows an example of the described reflection detection and selection with $k = 10$ loudest reflections. In addition to the peak and RMS amplitude values of each selected reflection the plot at the top shows the masking threshold [18] adapted to the RIR. To better estimate the audibility of reflections that arrive significantly later after the direct sound, the masking threshold could be further adjusted iteratively as a function of time [12]. However the current implementation seems to be a simple measure to validate whether the detected reflections (primarily the first reflections after the direct sound) are relevant. In the example shown in Fig. 2 it is interesting to see that all selected reflections are before the mixing time (MT = 58 ms). This indicates that at least in this example, the mixing time is a reasonably good predictor for the transition from a directional to diffuse sound field.

The plot at the bottom of Fig. 2 additionally shows the weighting function (WF) derived from the selected reflections, which is used in the BRIR synthesis (see Sec. 1.4) to decompose the RIR into a directional and diffuse part. To build the WF the envelope of the pressure response $p$ is constructed by convolving its absolute value with a Hann window of 3 ms. This envelope function is set to 1 for the windows of 1.5 ms around the TOAs of the direct sound and selected reflections. Finally, the function constructed in this way is smoothed with a 1-ms window to avoid too-strong edges, which leads to the WF shown in Fig. 2 (bottom) as an example.

### 1.1.3 Reflection Filters

In addition to TOA, DOA, and amplitude, early reflections can be described more precisely by their spectral characteristics. Parametric rendering that, unlike the presented BRIR synthesis, does not use signal components of the RIR (e.g., [4, 19, 12]) requires so-called reflection filters to adjust the magnitude spectra of (synthetic) early reflections according to the frequency-dependent absorption properties of the reflecting surfaces. Gaining broadband reflection filters from a monaural RIR, however, can be problematic because windows of sufficient length (e.g., a window of 10 ms is required to analyze frequencies down to 100 Hz) around a reflection often contain other reflections that arrive later. The result is an inaccurate reflection filter with a comb-filter magnitude response consisting of parts of multiple reflections.

To still derive useful reflection filters, the proposed encoding determines the magnitude response of each selected reflection in three asymmetric windows of different sizes. The first window is equal to the 1.5-ms direct sound window as described in Sec. 1.1.1 and thus can analyze frequencies down to about 667 Hz. The longest third window has a length defined by the predefined lowest frequency to be resolved, e.g., 10 ms for a boundary frequency of 100 Hz. The length of the second window is then calculated to have a boundary frequency between the boundary frequencies of windows one and three, i.e., in the given example the second window would have a length of about 2.6 ms, leading to a boundary frequency of about 384 Hz.

The actual magnitude response is then composed of the different frequency components that can be resolved by the respective window. By this, most reflections are described correctly, at least in the higher frequency range (f > 667 Hz for the 1.5 ms window), whereas in lower frequency ranges, comb-filter structures can occur again, since these frequency components are based on longer windows.

To compensate for the spectral influence of the sound source, the magnitude responses of the reflections are filtered with the inverse magnitude response of the direct sound, which is also constructed from the frequency com-

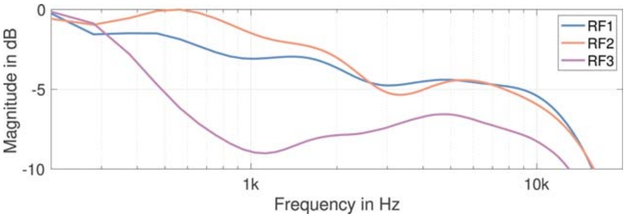

Fig. 3. Magnitude response $H_{ref_n}$ of the estimated reflection filters for the first three loudest reflections.

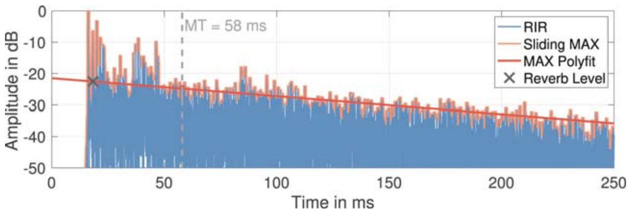

Fig. 4. Estimation of the reverberation level $L_{Rev}$ (gray cross), defined as the amplitude value of the decay curve (red line) at the TOA of the first selected early reflection.

ponents of the same three different windows. The procedure of inverse filtering is quite similar to the estimation of absorption coefficients from in situ measurements [20] and provides reflection filters that can be used for parametric rendering independent of the sound source used for the RIR measurement.

The reflection filters determined this way are nevertheless only a rough approximation of the absorption properties of the reflecting surfaces and become more and more inaccurate as the RIR progresses in time and reflection density increases. To compensate for those inaccuracies to a certain degree, the reflection filters are octave-smoothed in a final step. Fig. 3 shows the magnitude response $H_{ref_n}$ of three reflection filters as an example, obtained with the described method from the first three reflections ($n = \{1, 2, 3\}$) of the RIR shown in Fig. 2. Depending on the application, $H_{ref_n}$ can be transformed by frequency sampling into linear or minimum-phase FIR filters, or for example, with the Yule-Walker method into IIR filter coefficients [21].

### *1.1.4 Reverberation Level*

The last encoding step consists of the estimation of the reverberation level $L_{Rev}$. It describes the level of the diffuse sound field at the TOA of the first selected early reflection in the early directional part of the RIR. Reverberation level is an important parameter when synthesizing BRIRs using both a directional and diffuse reverberation component, since they must be combined at a certain level ratio (see, e.g., [4, 22]) to preserve the correct energy of the RIR.

In principle the DRR also describes such a level ratio, but since the energy is integrated over the entire length of the impulse response when calculating the DRR, the result depends on the length of the impulse response and thus also on the reverberation time. The proposed reverberation level parameter is independent of the reverberation time, so it is more convenient to use in practice, such as generating BRIRs with different reverberation times but similar level ratios between the directional and diffuse components. Furthermore the parameter can be used to estimate the diffuse energy in the early directional part or the amplitude at the mixing time for designing linear or cosine-squared ramps to fade in the diffuse reverberation (see, e.g., [19, 22]).

Fig. 4 shows an example of the calculation of the reverberation level. First the envelope of the absolute pressure response $|p|$ is estimated by calculating the maximum in a sliding window of 1 ms. The decay curve in decibels is then approximated by a first-order polynomial fit to the envelope in the time range between two and three times the mixing time. The selected time range ensures that there are no early reflections that distort the envelope, which would bias the determined decay curve. Lastly the decay curve is linearly extrapolated to obtain values for the time range before two and after three times the mixing time. The reverberation level is then defined as the amplitude value of the decay curve at the TOA of the first selected early reflection, which in this example is about $-22.5$ dB. Alternatively the amplitude of the reverberation at the mixing time can be determined by evaluating the decay curve at that time.

## 1.2 Spatialization

For spatialization, i.e., for assigning DOAs to the direct sound ($\angle_{ds}$) and the $k$ selected early reflections ($\angle_{ref_k}$), the proposed method offers three different possibilities. The following sections explain the three approaches in detail.

### *1.2.1 Pseudo-Randomized DOAs*

The first approach is inspired by the method introduced by Pörschmann et al. [11]. To obtain a correct representation of the direct sound its DOA must be known. The DOAs for the selected early reflections are based on a pseudo-randomized directional distribution stored in a lookup table. This lookup table with the DOAs can be derived, for example, from a shoebox-shaped room with non-symmetrically arranged source-receiver positions.

It is evident that the DOAs assigned this way match the actual reflection pattern only to varying degrees, depending on the situation. However, listening experiments showed that synthetic BRIRs with pseudo-randomized DOAs show quite high perceptual similarity with measured reference BRIRs [11, 23]. Thus, depending on the application, and if no further information about the sound field or space is available, such spatialization may be sufficient. Nevertheless the listening experiments by Pörschmann et al. [11] also showed that spatial attributes, such as apparent source width, localizability of the sound source, or listener envelopment, are significantly impaired by assigning incorrect DOAs. Thus more accurate spatialization methods should be preferred when possible.

### *1.2.2 Image Source Model*

If the approximate dimensions of the room (length, width, height) and the source and receiver positions are

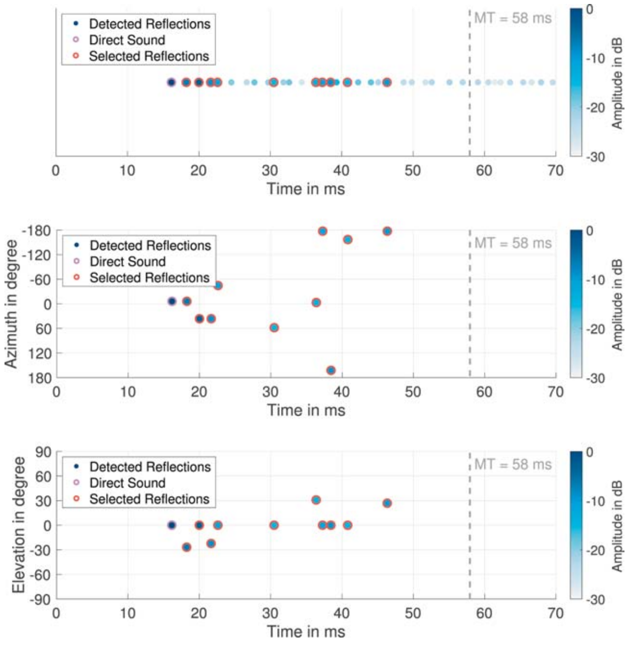

Fig. 5. Result of the spatialization based on a second-order image source simulation. Amplitude and TOA of the direct sound and the early reflections, acquired from the monaural RIR with the presented reflection detection (top). Spatialized direct sound and the selected $k = 10$ loudest reflections with assigned DOAs described by azimuth (middle) and elevation (bottom).

known, more appropriate DOAs for the selected early reflections can be obtained using an image source simulation for a shoebox-shaped room, which approximates the actual geometry of the room. The order of the image source simulation is freely selectable but since perceptual evaluations of parametric rendering have shown that six dynamically reproduced early reflections (theoretically the first-order reflections in a shoebox room) are already sufficient [19, 12], a second-order simulation is usually enough. In this spatialization mode, the DOA and distance of the direct sound can either be predefined or obtained from the simulation.

To estimate the DOAs of the selected early reflections, the TOAs estimated with the reflection detection (see SEC. 1.1.2) are compared with the TOAs calculated with the simulation, and reflections with equal TOA or smallest TOA differences are paired. The selected early reflections are then assigned the DOAs of the corresponding simulated reflections. The algorithm can be set first to assign the more important first-order reflections and then, in a second step, the second-order reflections to the remaining estimates. The described approach allows a relatively simple but much more correct spatialization than using pseudo-randomized DOAs, provided that the room's geometric information is available or can be determined.

Fig. 5 shows an example of the spatialization based on a second-order image source simulation. The plot at the top shows the TOA and amplitude of the direct sound and all detected early reflections, as extracted with the proposed encoding (see SECS. 1.1.1 and 1.1.2) from the RIR presented in Fig. 2. The two other plots show the azimuth (middle plot) and elevation (bottom plot) of the spatialized direct sound and the $k = 10$ loudest early reflections with assigned DOAs that are dynamically updated during rendering.

### 1.2.3 Precomputed DOAs

A precomputed DOA pattern can also be used for spatialization, making use of multichannel RIRs to derive DOA estimates. For example DOAs estimated in the context of SDM with open microphone arrays exploiting time differences of arrival (TDOAs) or with B-format array using pseudo intensity vectors (PIVs) [3, 24, 1, 25] can be passed to the spatialization. The direct sound and selected reflections are then assigned the DOAs listed according to the TOAs in the passed DOA vector. In the best case the DOA vector is post-processed, i.e., smoothed and with stabilized direct sound [26], to improve the perceptual quality of binaural renderings.

Using precomputed DOAs is an efficient way of combining an accurate spatialization with the scalable encoding and decoding of the presented method. The parametrization enables efficient storage and rendering of a multichannel RIR and a handful of parameters. Additionally, with only one source-receiver combination, the results can be extrapolated to other positions.

## 1.3 Extrapolation

The monaural and spatial parameters are sufficient for 3-DoF rendering at the receiver position, i.e., dynamic spatial audio reproduction for any head orientation of the listener (yaw, pitch, roll), for example using dynamic binaural synthesis. In this case only the DOAs of the direct sound and the early reflections ($\angle_{\mathrm{ds}}$ and $\angle_{\mathrm{ref}}$) have to be dynamically adjusted according to the head orientation by applying a corresponding rotation matrix. However for 6-DoF rendering, i.e., when the listener moves through the room, the amplitude and the TOA of the direct sound and the early reflections ($a_{\mathrm{ds}}$, $a_{\mathrm{ref}}$, $t_{\mathrm{ds}}$, and $t_{\mathrm{ref}}$) must also be adjusted accordingly. Since there are no further measurement points in the present case we refer to this as an extrapolation of the parameters to the new listener position.

The extrapolation adjusts the amplitudes, TOAs, and DOAs according to an image source model, as shown in Fig. 6, exemplified by direct sound and two reflections. For this purpose the TOAs are transformed into distance values describing the distance to the sound source (direct sound) or the so constructed image sources (reflections), and in combination with the DOAs, they are converted into Cartesian coordinates. The extrapolated DOAs and TOAs are then obtained by subtracting the displacement vector, describing the listener translation, and transforming the result back to spherical coordinates. The amplitudes are adjusted according to the change in distance to the sound or image source using the inverse-square law.

The extrapolated set of parameters describes the directional sound field for frontal head orientation and can again be adjusted according to the listener's head orientation by rotating the DOA vector. As the level of the diffuse reverberation is kept constant in rendering or BRIR synthesis,

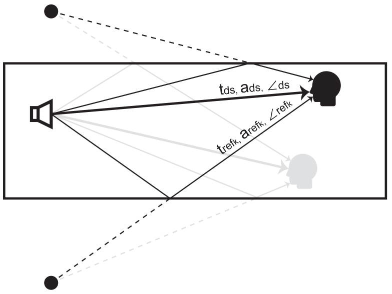
$t_{ds}$, $a_{ds}$, $\angle_{ds}$
$t_{refk}$, $a_{refk}$, $\angle_{refk}$

Fig. 6. Extrapolation to a new listener position (gray to black) by adjusting the amplitudes, TOAs, and DOAs of the direct sound and the (here shown as example two) early reflections based on an image source model.

adjusting the amplitudes of the directional components and filtering the direct sound according to the source directivity also changes the DRR appropriately as the listener moves through the room (see Secs. 1.4 and 2). Thus two important cues for distance perception in enclosed spaces, level and DRR [27], are correctly reproduced.

## 1.4 BRIR Synthesis

The parametrization, spatialization, and optional extrapolation provide sufficient information for parametric BRIR synthesis or real-time parametric rendering. Thus completely synthetic BRIRs (i.e., BRIRs without any signal components of the original RIR) could be generated based only on the parameters using one of the many synthesis algorithms (see, e.g., [28, 29, 4, 22, 12]). However the synthesis that we have implemented as part of the *Paraspax* framework uses signal components of the monaural RIR in addition to the parameters to generate the BRIRs, with the great advantage that the essential sound characteristics of the room are preserved.

The diffuse components (the binaural diffuse reverberation) and directional components (the spatialized direct sound and early reflections) are first synthesized separately and finally added to form a BRIR. Besides the monaural RIR and parametric description (as parts of the monaural RIR are used for the synthesis, only the amplitudes, TOAs, and DOAs determined in encoding are required), the synthesis needs a full-spherical set of Head-Related Transfer Functions (HRTFs, i.e., the frequency-domain equivalent of the HRIR) and, if the directivity of the sound source is to be considered, corresponding directivity measurements. As part of a 6-DoF spatial audio system the synthesis can be used to precompute BRIRs for arbitrary head orientations and room positions, which can then be used for real-time convolution (see Sec. 3).

### 1.4.1 Diffuse Components

The binaural diffuse reverberation is generated based on the monaural RIR by convolution diffusion [1], i.e., the monaural RIR is filtered with binaural noise, which is quite similar to filtering with rectangular noise bursts in order to decorrelate omnidirectional reverberation for loudspeaker representation [30]. Fig. 7 illustrates the method implemented in *Paraspax*, which is essentially based on the convolution diffusion approach proposed by Pörschmann et al. [11]. First binaural noise is generated, which is two-channel white noise, filtered with the diffuse field response (also called common transfer function [31, ch. 1]) of the applied HRTF set and a coherence filter modeling diffuse-field interaural coherence (IC) [32]. The binaural diffuse reverberation is then synthesized by convolving small chunks of the RIR (0.67 ms; 32 samples at 48 kHz sampling rate) with chunks of the binaural noise (2.67 ms; 128 samples at 48 kHz sampling rate) and summing with overlap-add. This processing results in binaural reverberation with the frequency-dependent decay of the monaural RIR as well as with the diffuse-field IC and the associated spaciousness.

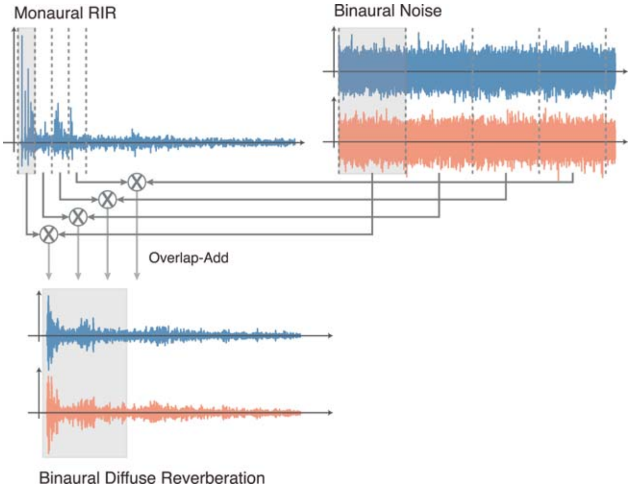

Fig. 7. Synthesis of the binaural diffuse reverberation by convolution diffusion. Small chunks of the RIR are convolved with small chunks of binaural noise and summed with overlap-add (adapted from [11]).

We determined the time constants for the RIR and the noise chunks with informal listening comparing to measured references. The listening revealed that longer noise blocks lead to a more spatial and less neutral-sounding reverberation, as also informally determined in a previous study [1]. Our informal evaluation showed that a block length of 32 samples for the RIR and 128 or 256 samples for the noise (at a sampling rate of 48 kHz) provides the best results.

### 1.4.2 Room Impulse Response Decomposition

Directional components mainly characterize the early part of a BRIR, where most energy is contributed from specular reflections. As the BRIR progresses in time, the relative contribution of diffuse energy progressively increases [33, ch. 4]. Listening experiments showed that mixing directional and diffuse components in the early part of a BRIR provides higher perceptual quality than BRIRs with purely directional components, indicating that synthetic BRIRs should exhibit diffuse sound also in the early part [4, 19].

The approach presented here synthesizes the early part of a BRIR as the superposition of weighted specular components and diffuse components. Fig. 8 (top) illustrates

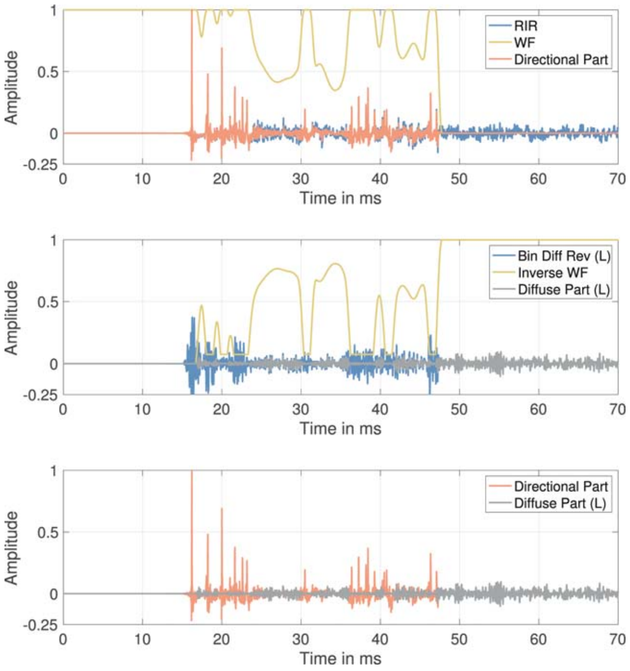

Fig. 8. RIR decomposition in the early part. Extraction of the directional (specular) components by applying the weighting function (top). Estimation of the diffuse components by applying the inverse weighting function to the synthesized binaural diffuse reverberation (middle). For illustration, superimposed monaural specular and (left channel) binaural diffuse part (bottom).

how the monaural RIR is decomposed to extract the directional (specular) components applying the weighting function (WF) estimated with the reflection detection (see Sec. 1.1.2). The extracted directional part is the basis for the synthesis of early reflections in the synthetic BRIRs (see Sec. 1.4.3).

Fig. 8 (middle) shows how the diffuse component in the early part is estimated by applying the inverse WF to the binaural reverberation—only the left channel is shown in the example. The inverse WF is constructed as the square root of the inverted weighting function, limited to the reverberation level estimated in encoding so that diffuse energy is also preserved in specular events (i.e., where WF is equal to 1). To illustrate the decomposition, Fig. 8 (bottom) finally shows the extracted monaural specular part overlayed with the left channel of the binaural diffuse part.

### 1.4.3 Directional Components

The direct sound and selected early reflections are the only components adjusted according to the listener's head orientation or position in the *Paraspax* synthesis, which allows generating BRIRs for 6-DoF spatial audio reproduction. The synthesis can be performed for any number of head orientations as well as for any listener position. The head orientations are represented by a spatial sampling grid with azimuth and elevation describing either the relative orientation to the sound source (global coordinate system) or the listener-related head orientation (local coordinate system), whereas the listener position is described by Cartesian coordinates in a global coordinate system with the origin being the measurement position of the monaural RIR.

The *Paraspax* algorithm performs the synthesis of the directional components in the following steps. First, chunks of the RIR are extracted in the non-symmetric 1.5-ms windows around the TOAs of the direct sound and the selected early reflections. When an extrapolated position is synthesized the chunks are adjusted in amplitude with the amplitude factors derived from the previous parameter extrapolation (see Sec. 1.3). If sound source directivity measurements are available a directivity filter is applied to the direct sound in the next step.

Since the sound source directivity is usually already imprinted in the direct sound extracted from the monaural RIR, the direct sound is filtered only according to the change of the directivity corresponding to a change of position relative to the sound source. Thus, for the synthesis at the measurement position, the directivity filter has a flat magnitude response over the entire frequency range (i.e., no filtering). However for extrapolated positions, the direct sound is filtered according to the direction and frequency-dependent change in sound source directivity. In the present case the directivity measurements are stored as spherical harmonics (SH) coefficients at a sufficiently high spatial order $N \geq 35$. Using the SH description allows artifact-free SH interpolation of the directivity to obtain suitable directivity filters for any radiation direction [34].

Next the (processed) RIR chunks are convolved with HRTFs according to the DOAs for the respective head orientation, resulting in a specific directional pattern for each point of the spatial sampling grid. To calculate the DOAs first estimated for frontal head orientation at the measurement or extrapolated position, a rotation matrix according to the head orientation is applied (see, e.g., [26]). As with the directivity measurements, the HRTF set is stored as SH coefficients at $N \geq 35$, allowing HRTFs to be obtained for any direction.

Next the resulting directional components (i.e., the convolution result of RIR chunks and HRTFs) are placed at their respective position in time according to the estimated or extrapolated TOAs. In a last step, the amplitude of the synthetic direct sound for frontal head orientation at the measurement position (azimuth $\phi = 0°$ and elevation $\theta = 0°$ relative to the source) is compared with the amplitude of the direct sound from the monaural RIR, and the level of the entire synthetic directional component is adjusted accordingly. This level adjustment, which is a constant factor applied to all synthesized directional components, compensates for level changes that may occur, for example, due to convolution with non-level normalized HRTFs, which would drastically change the DRR in the final BRIRs.

Fig. 9 shows as an example the result of the described synthesis of the directional components of a synthetic BRIR for the measurement position and frontal head orientation. The synthesis is based on the extracted direct sound and extracted ten loudest reflections of the monaural RIR shown in Figs. 2 and 8.

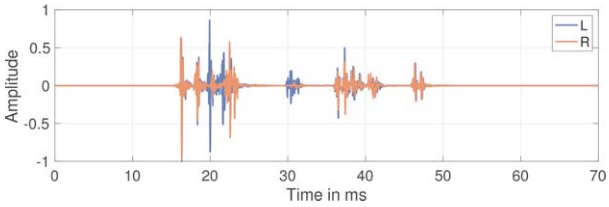

Fig. 9. Synthesized directional components.

### 1.4.4 Composition of Directional and Diffuse Components

To generate the final synthetic BRIRs, the directional components, which are two-channel BRIRs for each head orientation at each listener position, are combined with the diffuse components for the early part and with the binaural diffuse reverberation characterizing the late part of the BRIR. Fig. 10 (top left) shows the early part of the synthetic BRIR, which is a composition of the weighted synthetic binaural reverberation (see SEC. 1.4.2) and the directional components synthesized based on the specular part of the monaural RIR (see SEC. 1.4.3). The top right plot in Fig. 10 shows a longer segment of the same synthetic BRIR with a logarithmic amplitude scale to better illustrate the decay of the binaural reverberation.

As an example, the two plots at the bottom left and right in Fig. 10 show a synthetic BRIR for an extrapolated position (two meters back and one meter to the right from the measurement position) with the same head orientation. The amplitude and TOA shifts are clearly visible in the early part, whereas the late part is the same for all BRIRs.

The *Paraspax* toolbox offers various post-processing options to further filter the synthetic BRIRs. Hence filters can be applied to compensate at least to some extent for room modes that appear in the monaural RIR, to change the IC of the directional or diffuse components, or to high-pass the BRIRs or only the diffuse components to control low-frequency ($f \leq 50$ Hz) spaciousness and reverberation.

## 2 PHYSICAL EVALUATION

For the physical evaluation of the *Paraspax* synthesis we compared measured RIRs and BRIRs with synthesized BRIRs concerning various room acoustic monaural and binaural parameters. Even if the synthesis method does not attempt to physically reconstruct the room response, synthetic and measured impulse responses should ideally match within the limits of the respective just-noticeable differences (JNDs) of room acoustic parameters to achieve satisfactory perceptual results.

### 2.1 Room and Measurement Setup

We conducted measurements and implemented a 6-DoF spatial audio system (see SEC. 3) based on BRIRs synthesized with the *Paraspax* toolbox in a mostly empty shoebox room. Fig. 11 (top) shows a cross-section of the room with dimensions 11.73 m × 4.74 m × 4.62 m (length × width × height). We measured the impulse responses using the swept-sine technique with an Earthworks M30 as well as with a KEMAR dummy head on a 4 m × 3 m rectangular grid of 1-m resolution and a measurement height of 1.40 m. It is important to note that we measured the KEMAR BRIRs with Brüel & Kjær 4101 binaural in-ear microphones (placed in the ear canal of the KEMAR) and not with the microphones integrated in the dummy head. The sound sources were four Genelec 8020 loudspeakers at different positions in the room and different heights.

Fig. 11 (bottom) shows a picture of the room with the KEMAR dummy head (wearing AKG K1000 headphones) at position 10 of the grid, the 4 Genelec loudspeakers, and an optical tracking system (OptiTrack) covering the entire room. The room has plain walls, a large glass window, and a concrete floor, leading to distinct early reflections and a relatively high reverberation time of $RT_{30} = 0.9$ s (average between 0.25 and 8 kHz).

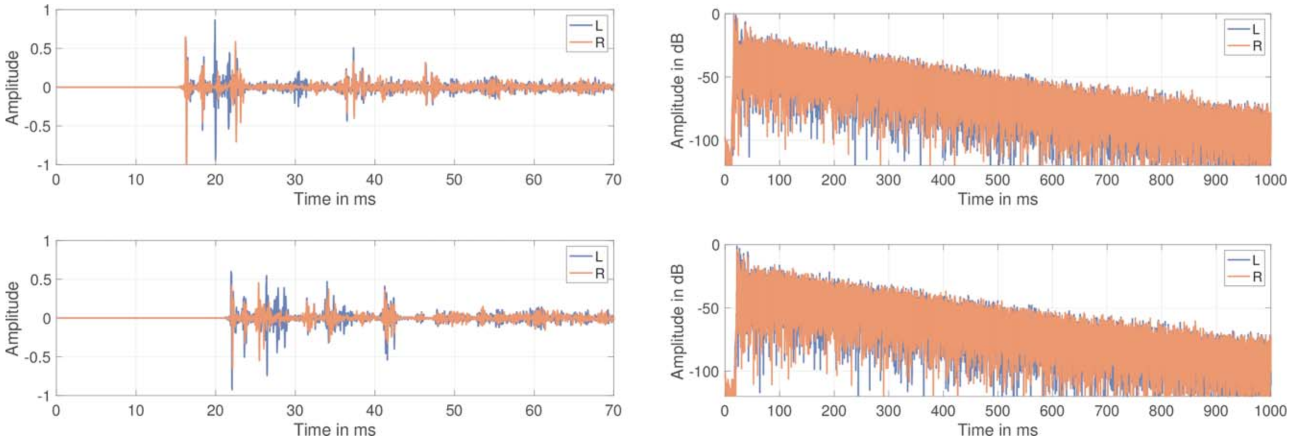

Fig. 10. Early part (left) and longer segment (right) of synthesized BRIRs, composed of directional components and binaural diffuse reverberation. The plots at the top show a synthesized BRIR for the measurement position of the monaural RIR; the plots at the bottom a BRIR for an extrapolated position (two meters back and one meter to the right from the measurement position).

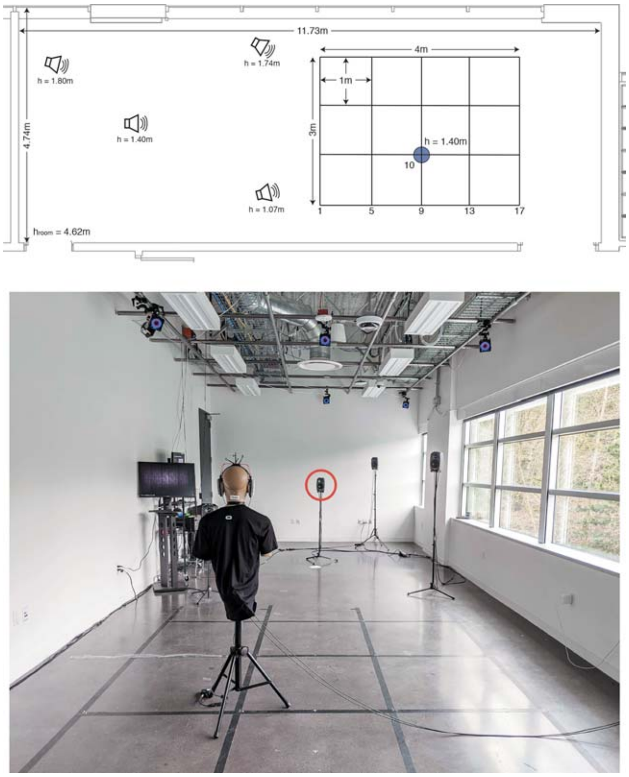

Fig. 11. Cross-section of the room with the positions of the four loudspeakers and the measurement grid/listening area (top). Picture of the room with a KEMAR dummy head at position 10 of the grid, the four Genelec 8020 loudspeakers, and the OptiTrack system (bottom).

## 2.2 Synthesized BRIRs

We have chosen position 10 as the reference position for this study, which means that all synthesized BRIRs for each position, head orientation, and loudspeaker for this room are based on 1 monaural RIR for each loudspeaker measured at position 10. Due to the high number of measurement points we limited the physical evaluation presented here to BRIRs synthesized for loudspeaker 2 (highlighted loudspeaker in Fig. 11 (bottom)), which has a height of 1.40 m and distance to position 10 of about 5.5 m. For comparison we synthesized BRIRs for frontal head orientation (local head-related coordinate system) based on the monaural RIR measured at position 10 with loudspeaker 2 as the source, with the spatialization based on a second-order image source simulation (see SEC. 1.2). The loudspeaker was slightly offset to the right in azimuth from position 10 (approximately 6°).

For the synthesis we used KEMAR HRTFs, measured with Knowles FG-23329 miniature microphones at the blocked ear canal of the dummy head. The measurements were done on a high-resolution spherical sampling grid (9,720 directions) in the anechoic chamber at Facebook Reality Labs Research. The HRTFs have been low-frequency corrected below 200 Hz [35] and transformed to SH domain at $N = 35$ using the discrete spherical Fourier transform with Tikhonov regularization [34]. The directivity data for the Genelec 8020 loudspeaker were taken from the BRAS database [36] and also transformed to the SH domain at $N = 35$.

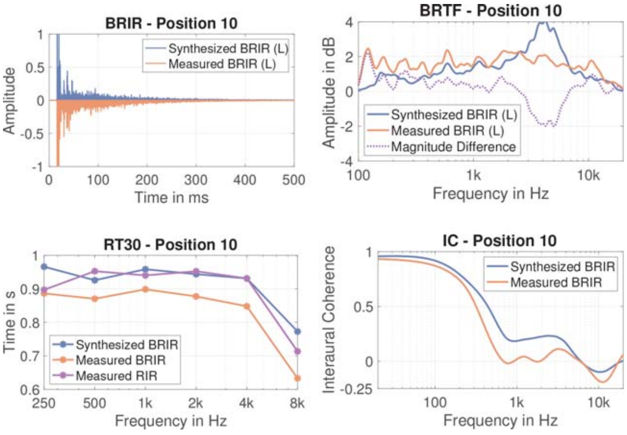

Fig. 12. Comparison of measured and synthesized impulse responses at position 10 in terms of time-energy structure (top left), magnitude response (top right), reverberation time $RT_{30}$ (bottom left), and frequency-dependent IC (bottom right).

### 2.2.1 Measurement Position

Fig. 12 provides a first comparison of the measured and synthetic impulse responses at position 10. The synthesized and measured broadband pressure BRIRs for the left ear (top-left plot) show good similarity in their overall time-energy structure with matching amplitude and time events. However the measured BRIR shows a finer resolved early part, mainly because the synthesis is based only on the ten loudest reflections.

The left-ear binaural room transfer functions (BRTFs) in the top-right plot show a good agreement between 0.2 and 2 kHz with magnitude differences of 1 dB or less. However in the frequency range between 2 and 5 kHz, there are noticeable deviations in magnitude of about 2 dB. These variations may occur because the BRIRs were measured in a different way than the HRTFs used for synthesis, i.e., the BRIRs were measured with microphones in the ear canals (not fully blocked) and the HRTFs were measured with different microphones on the blocked ear canal. Furthermore the diffuse field response obtained from the employed HRTF set and applied to the synthesized binaural reverberation (see SEC. 1.4.1) may increase slightly too much between 3–6 kHz.

The plot at the bottom left of Fig. 12 compares the reverberation times of the different impulse responses, where $RT_{30}$ of the BRIRs was calculated as the mean value of $RT_{30}$ for the left and right ear. The plot reveals that the synthesized BRIR and measured RIR have nearly identical frequency-dependent reverberation times, indicating that the proposed RIR decomposition and reconstruction when synthesizing BRIRs works well. Furthermore it shows that the presented convolution diffusion approach is well suited for synthesizing binaural reverberation from a monaural RIR while maintaining the decay of it.

The reverberation time estimated from the measured BRIR is slightly lower, possibly due to the dummy head acting as a directional receiver. To further investigate this we estimated the reverberation time for BRIRs measured

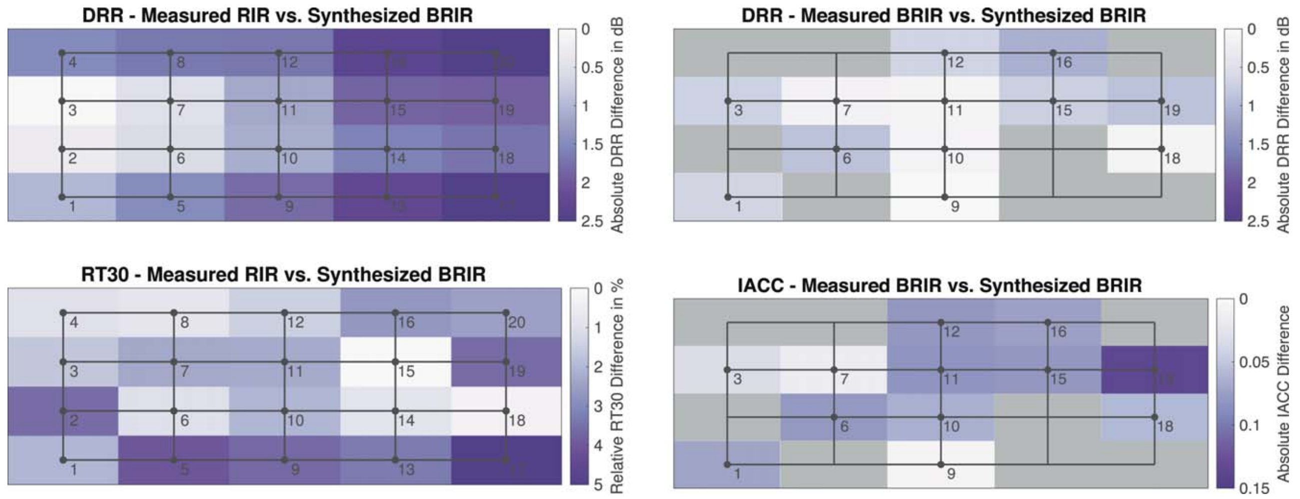

Fig. 13. Differences in DRR (top left and right), RT$_{30}$ (bottom left), and IACC (bottom right) between measured and synthesized impulse responses at the grid positions. The colored rectangular fields visualize the magnitude of the differences at the grid positions, which are represented by a small dot in the center of the respective rectangular field. The gray fields (top and bottom right) denote positions without measurement data.

with different head orientations, confirming that the reverberation time varies slightly depending on the head orientation. Overall the reverberation times estimated from the measured BRIRs (for frontal head orientation) are lower at all positions in the room (see Table 1 in the APPENDIX). However as the synthetic BRIR is based on the measured RIR it is sensible to expect that in the present evaluation the reverberation times of these two impulse responses are mostly identical. Further studies are required to examine the effect of head orientation on estimated reverberation time, as well as the audibility of these deviations.

Lastly the bottom-right plot shows the frequency-dependent IC of the measured and synthesized BRIRs [28]. The plot reveals a very similar IC over frequency, although the synthesized BRIR has a slightly higher IC than the measured BRIR, especially above 200 Hz. These deviations indicate that the measured BRIR has more incoherent directional events than the synthesized BRIR.

### 2.2.2 Extrapolated Positions

To further examine the performance of the presented 6-DoF synthesis we compared DRR, RT$_{30}$, and Interaural Cross-Correlation (IACC) of synthesized and measured impulse responses at different points on the measurement grid. As mentioned above, all synthetic BRIRs were based on the monaural RIR measured at position 10 and were generated for frontal head orientation (local head-related coordinate system) by applying the extrapolation and synthesis described in SEC. 1. Accordingly the KEMAR BRIRs were also measured for frontal head orientation at the different positions in the room (see also Fig. 11).

The DRR was calculated with a direct sound window of 1.5 ms and for two-channel BRIRs, the direct and diffuse sound energy of the left and right channels were summed before calculating the DRR [37]. The broadband RT$_{30}$ was calculated as the mean value of the reverberation times in the octave bands from 0.25–8 kHz. In the case of BRIRs the reverberation times determined per channel were averaged to obtain a single value.

Table 1 in the APPENDIX lists the individual DRR, RT$_{30}$, and IACC values for the respective impulse responses and grid positions. For better comparability off the estimated values, the plots in Fig. 13 show the deviations in DRR, RT$_{30}$, and IACC between measured and synthesized impulse responses at the grid positions. The colored rectangular fields visualize the magnitude of the differences at the measurement positions, with each measurement position represented by a small dot in the center of the respective rectangular field.

The absolute DRR differences between measured RIRs and synthesized BRIRs (top-left plot) exceed 2 dB only at a few positions (points 13, 16, 17, and 20), with a maximum of about 2.6 dB at position 20. Larsen et al. [38] determined JNDs of about 2–3 dB in rooms with a DRR of 0 or 10 dB and JNDs of about 6–8 dB in rooms with a DRR of −10 or 20 dB. In the present case the DRR ranged between −2.5 to −11 dB for the monaural RIRs, and thus the absolute DRR differences are well below estimated JNDs for DRR changes.

However as there are generally significant differences between the DRR estimated from a BRIR or from an omnidirectional RIR, that is, due to the directivity of the artificial head or due to the influence of the source angle [37], we have also compared the DRR of synthesized and measured BRIRs at different positions in the room. The results presented in Fig. 13 (top right) reveal absolute DRR differences far below JND, with a maximum difference of only about 1 dB at position 16 and differences of even less than 1 dB at all other evaluated positions. Please note that we have only made KEMAR measurements at those positions where numbers appear in the plot and have grayed out the rectangular fields without data.

The bottom-left plot in Fig. 13 shows the relative RT$_{30}$ difference between measured RIRs and synthesized BRIRs at all 20 positions. The largest deviation occurs at position 17, with about 5% relative difference between the broadband reverberation times of the synthesized BRIR and monaural reference RIR. At all other positions the relative difference is even below 5% and thus clearly below the JND for reverberation, which ISO 3382-1 lists as 5% according to Seraphim [39]. When considering the $RT_{30}$ differences, however, it should be noted that all BRIRs use the same binaural reverberation synthesized from the monaural RIR measured at position 10, and therefore the reverberation time of the BRIRs is nearly identical at all points in the room. Thus the analysis presented here rather gives information about how homogeneous the sound field is in the room and how well the criterion of diffusity, which we indirectly apply in the synthesis, is fulfilled.

Lastly, Fig. 13 (bottom right) shows the absolute IACC differences between synthesized and measured BRIRs at 12 positions in the room. The differences exceed the JND for IACC, which is defined as 0.075 according to ISO 3382-1, only at position 19. At positions 6, 11, 12, and 15 the differences are in the JND range, whereas at all other positions the differences are clearly below the JND. Overall the broadband IACC for the KEMAR BRIRs ranges from 0.16 to 0.40 and from 0.14 to 0.43 for the synthesized BRIRs.

Interestingly the broadband IACC of the synthesized BRIRs is often slightly higher than that of the measured BRIRs, which is in line with the observations for the frequency-dependent IC at position 10 (see Fig. 12). These findings indicate that the measured BRIRs contain more directional and incoherent components (i.e., early lateral reflections). However as the differences exceed the JND for IACC only at one position we assume that the perceived spaciousness is similar throughout the room, regardless of whether synthetic or measured BRIRs are used for 6-DoF spatial audio reproduction.

## 3 SIX-DEGREES-OF-FREEDOM VIRTUAL ACOUSTIC ENVIRONMENT

The physical evaluation based on room acoustic parameters showed a good performance of the *Paraspax* method. However the study's main goal was implementing a perceptually plausible 6-DoF VAE based on one monaural RIR, which can be used, for example, for perceptual studies on AR audio. For this reason we implemented a 6-DoF framework using precomputed BRIRs synthesized with the *Paraspax* toolbox, with which demo applications and plausibility studies according to Lindau and Weinzierl [40] can be performed.

### 3.1 Real-Time Framework

The block diagram in Fig. 14 gives an overview of the 6-DoF framework, deployed in the room where we also performed the physical evaluation (see Fig. 11). A Matlab software developed for the framework takes over the central control. The software receives tracking data from an optical tracking system (OptiTrack with 7 Prime 41 cameras) installed in the room at an update rate of 150 Hz and control data from a Numark Orbit MIDI controller at an update rate of 30 Hz. The OptiTrack system provides the position of the four Genelec 8020 loudspeakers and the position and orientation of the listener, and the Matlab software calculates from these data the relative orientation of the listener to the loudspeakers (azimuth, elevation, and distance) and the listener's absolute position in the room (X, Y, Z).

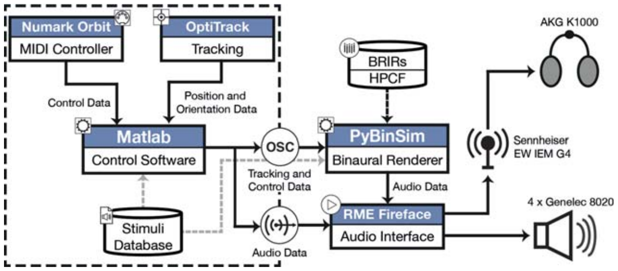

Fig. 14. Block diagram of the implemented 6 DoF real-time framework for demo applications and psychoacoustic experiments in AR audio.

In the demo application, the listener can use the wireless MIDI controller to switch between real and virtual sources, switch between 1 of the 4 (virtual) sources, switch randomly between 1 of the 50 loudness-normalized test stimuli, or stop or pause playback. These control options enable a detailed comparison between real and virtual sound sources while moving through the room. In the case of a plausibility study the participant uses the MIDI controller to enter the answers. The Matlab software sends tracking and control data via OSC to the binaural renderer and audio data to a RME Fireface UCX audio interface if a real source (i.e., a loudspeaker) is played.

As the binaural renderer we used PyBinSim, a Python tool for real-time dynamic binaural synthesis [41]. The renderer convolves the dry audio signal with BRIRs, which have to be precomputed according to the desired spatial resolution for each head orientation and position in the room. The renderer allows using BRIRs split into an early and late part. In this case, only the early (directional) part needs to be adjusted according to the listener's head orientation and position by switching the early BRIRs, whereas the late part, i.e., the binaural diffuse reverberation, which is the same in all BRIRs, is covered by a single reverberation BRIR. This approach reduces the required memory and improves performance.

The BRIRs for each of the 4 sources were all based on 1 monaural measurement at position 10. For encoding of the four monaural RIRs, we used the *Paraspax* method as described in Sec. 1, with the spatialization based on a second-order image source simulation (see Sec. 1.2). We synthesized BRIRs using the method described in Sec. 1.4 with a spatial resolution of 4° in azimuth and 10° in elevation (restricted to ± 50°), resulting in 990 BRIRs per position. The uniform grid resolution, that is, the extrapolation steps in X and Y direction, was 0.25 m, which leads to 285 points on the 4 m × 3 m listening area. Given the four different sources, a total of 1,128,600 (early) BRIRs were rendered.

To enable a direct comparison between real and virtual sound sources we used the extraaural AKG K1000 headphones for binaural playback, which were equalized with minimum phase FIR headphone filters designed with automatic regularization [42]. For better mobility the headphone signal was transmitted via a Sennheiser EW IEM G4 wireless in-ear monitor system. Furthermore the real sources (i.e., the loudspeakers) and their virtual representation over headphones were matched in loudness so that no level changes occurred when switching between real and virtual sources. The overall playback level in the listening area was $L_{\mathrm{Aeq}} = 60$–$70$ dB, measured for the closest sound source (i.e., loudspeaker 1).

The latency of the system was estimated to be approximately 66 ms, which is within the range of empirically determined thresholds of just detectable latency of 60–75 ms [43], and thus should be low enough in most real-life cases. The system latency consists of approximately 41 ms latency by the OptiTrack system [44], approximately 7 ms latency by processing the tracking data in Matlab (update rate of 150 Hz), approximately 6 ms due to OSC latency [45], about 11 ms due to the 512 samples buffer size in PyBinSim ($f_s = 48$ kHz), and additionally about 1 ms due to the RME Fireface UCX running with 64 samples buffer size ($f_s = 48$ kHz).

For a similar system (same tracking and playback system but a different binaural renderer), a similar motion to sound latency of about 70 ms was measured [46], suggesting that the latency estimate for the present 6-DoF VAE is quite accurate. However some participants in the plausibility experiment (see SEC. 3.2) could perceive head-tracking latency during high-speed rotations, suggesting that the actual latency of the system was higher than estimated. Accordingly latency measurements and possibly optimization of the system to reduce latency is inevitable in future work.

## 3.2 Plausibility Experiment

To assess the plausibility of the presented 6 DoF parametric spatial audio system we implemented a two-alternative forced-choice (2AFC) experiment as proposed by Lindau and Weinzierl [40]. During each trial a real or virtual source was randomly presented to the participants and the participant had to answer whether it was a real sound source (i.e., the sound came from one of the four loudspeakers) or a virtual sound source (i.e., the sound was reproduced over headphones).

The answers were analyzed with signal detection theory (SDT), i.e., the percentages of correct answers resulting from the 2AFC test were transformed to the criterion-free measure for the sensory difference $d'$ between the presentation of a real or virtual sound source. A $d'$ of 0 (50% correct answers; guessing rate) indicates inaudible differences, i.e., the VAE is perfectly plausible, whereas $d' > 0$ suggests audible differences between the presentations. A common critical value above which an auralization is considered to be no longer plausible is $d' = 0.82$, which corresponds to a 2AFC detection rate of $p_{2\mathrm{AFC}} = 0.72$ [47].

Each run had 120 trials consisting of 30 different test signals and 4 sound source positions. Thus, each of the 30 test signals was played exactly once with each of the 4 sources, whereby half of the trials were real and half were virtual sound sources. The order of test signal, sound source, and real or virtual presentation was randomized for each participant. The test signals were loudness normalized monophonic audio content with a length of 10 s. It included white and pink noise bursts, male and female speech, vocals, wind, string, percussion instruments, and synthesizer sequences.

The high number of different test signals prevents familiarization and the fact that the same combination of a test signal and sound source position is never presented both as a real and virtual source corresponds to the basic idea of plausibility testing [40, 47]. This procedure clearly distinguishes the plausibility test from a test on authenticity, where the same content is presented as a real and virtual source in short succession [48].

During the experiment participants were asked to walk along a predefined path in the listening area. According to the test paradigm, a random test signal was played from one of the four sources (real or virtual) and participants had to answer whether the presentation was real or virtual by pressing the corresponding button on the wireless MIDI controller. The participants were allowed to walk the path at their desired speed, turn around, or stop for a short time. In general, however, they were asked to cover the entire listening area by constantly walking along the path.

### 3.2.1 Results

Data collection was restricted due to the effects of the COVID-19 pandemic, resulting in a pilot study with four participants (1 female, 3 male, aged 25 to 35 years). All participants were highly experienced with binaural reproduction since they work full-time in this field and they were not naive as to the purpose of this study nor the technical background of the 6-DoF VAE.

The participants had a detection rate of $p_{2\mathrm{AFC}} = 0.61$, 0.77, 0.84, and 0.87, corresponding to $d' = 0.40$, 1.04, 1.41, and 1.59. Concerning the critical value of $d' = 0.82$, the presented 6-DoF parametric spatial audio system was thus perceptually plausible (in this strictly defined manner) for 1 participant, whereas the other participants could more reliably detect whether a real or virtual sound source was presented.

There are several reasons why the participants could distinguish between virtual and real sources, as revealed by informal discussions after the experiments. First of all it is important to mention again that the pilot study participants are to be classified as extremely critical listeners since they are experts in this field and have also followed the development of the system. As typical indicators for detecting a virtual source the participants mentioned the apparent source width, the listener envelopment, and slight localization inaccuracies between a virtual source and physically present loudspeaker, especially at front grid positions very close to speaker 1 or 4.

These observations indicate that the BRIR synthesis has to be further improved, especially in terms of spaciousness (IC and IACC), and that the spatial resolution of 4° in azimuth and 10° in elevation, which we have chosen as a compromise between accuracy and memory capacities, might be too low. The greater apparent source width of the virtual sources sometimes perceived by the participants as well as localization errors may also have been caused by the employed non-individual KEMAR HRTFs.

However, regardless of the quality of the synthesized BRIRs, there were several system-related effects that allowed participants to distinguish between a real and virtual source. Two of the four participants used head-tracking latency during high-speed rotations, which caused the virtual sound source to lag behind, to detect virtual sources. Furthermore, especially with periodic continuous test signals like, for example, legato strings, participants could perceive the BRIR switching at grid transitions, i.e., when walking through the room and passing various grid cells. These findings are in line with a study from Werner et al. [49], which showed that especially at grid transitions (close to the sound source), participants perceived abnormalities in the synthesis. The artifact was more pronounced with audio content such as strings since the different BRIRs are cross-faded at the grid transitions, which leads to audible phase shifts, especially with periodic content. Speech or drum test signals therefore did not lead to such audible artifacts at grid transitions, which is also in line with the results from Werner et al. [49].

Furthermore despite moderate playback levels, the participants could sometimes feel the headphones' vibration when a binaural signal was presented and could thus detect whether the source was real or virtual. This problem could easily be solved with other more extraaural headphones. Lastly we could observe an apparent training effect, that is, the participants got to know all these artifacts throughout the experiment and then recognized the virtual source often due to the artifacts but not necessarily due to audible differences to a real source.

## 4 DISCUSSION

Parametric rendering is a promising approach to create perceptually plausible spatial audio renderings of high quality for applications with low computing capacity, such as AR applications. Most parametric encoding methods rely on SRIR measurements, such as first-order Ambisonics (FOA) or open array measurements [1, 3, 2, 5, 6], or even higher-order rigid-sphere array measurement [4, 25]. Those methods usually auralize the sound field only at the measurement point for the listener's head orientation, i.e., the sound field is only rendered for 3 DoF.

More recently proposed methods for 6-DoF rendering typically require at least one [7, 8] or even multiple SRIRs [9, 10] from different positions in the room and apply extrapolation based on one SRIR or interpolation between the distributed SRIRs to incorporate listener translation. However 6-DoF rendering based on a parametric description of the sound field at one position in the space, encoded from a single RIR measurement, is rare. A perceptually plausible parametric 6-DoF rendering based on one RIR seems to be a good solution, though, e.g., for AR-glasses to augment a real environment with virtual sources, taking into account the listener's movements.

With the *Paraspax* method presented in this paper we have taken a first step toward scalable and perceptually motivated parametric 6-DoF rendering based on the encoding of a single monaural RIR. The three basic processing blocks for encoding include the parametrization of the monaural RIR to determine basic monaural room acoustic parameters and TOAs of the direct sound and early reflections, the spatialization to assign DOAs to the direct sound and the selected reflections, and the optional extrapolation for listener translation. The estimated parameters can then be used for synthesizing BRIRs based on the monaural RIR, as presented in this paper, or for real-time parametric rendering (see, e.g., [29, 50]). As the encoding provides the (perceptually) most relevant parameters to describe the sound field, the real-time rendering could be performed without any signal components of the monaural RIR using computationally more efficient auralization methods than the real-time convolution with a large number of precomputed BRIRs.

The presented method adapts some ideas from the implementation of Pörschmann et al. [11], who published a first approach to synthesize BRIRs that are generally suitable for 6-DoF rendering based on an omnidirectional RIR. In comparison, however, the *Paraspax* method provides a scalable and perceptually motivated reflection detection, i.e., the number of salient reflections [12] that are en and de-coded for rendering can be adjusted according to the available computing resources. Furthermore the three different options for spatialization allow estimating DOAs without any spatial information (pseudo-randomized), based on few geometric data (image source model), or based on SRIR measurements or simulations (precomputed). Thus even if only monaural parameters are available the *Paraspax* method allows 6-DoF parametric spatial audio rendering.

Of course spatial percepts such as apparent source width, localizability, or listener envelopment can be impaired by only partially correct early reflection DOAs, but listening experiments also showed a high perceptual similarity between measured reference BRIRs and synthesized BRIRs with pseudo-randomized DOAs [11, 23], suggesting that in cases where there is no geometric or spatial information at all, such an auralization may still be a sufficiently good solution. On the other hand passing precomputed DOAs allows combining the *Paraspax* en and de-coding with correctly estimated DOAs. Using image source simulation for spatialization provides DOA estimates for which the accuracy depends on the room geometry and which therefore become less accurate the more the actual room geometry deviates from the assumed shoebox geometry.

Informal listening for the present room showed that spatialization with pseudo-random DOAs produced audible differences, especially in terms of listener envelopment and spaciousness, compared to the other two spatialization methods (ISM and precomputed DOAs from SDM measurements), which could not be perceptually distinguished. However since the examined room is very close to the shape of a shoebox we did not expect significant perceptual differences between these two spatialization methods.

The *Paraspax* method also includes the sound source directivity in the BRIR synthesis. Especially due to the emerging research in 6-DoF audio reproduction, the topic of source directivity has gained more attention again [51, 52]. In particular scenarios where the listener can walk around a virtual sound source should benefit from considering the directivity, but even if the sources are outside the listening area, incorporating the directivity leads to better technical and perceptual results [51].

In the present plausibility experiment, however, participants could not walk around the loudspeaker, limiting our results concerning the benefit of including source directivity in the BRIR synthesis. Even though directivity effects are also audible when the loudspeakers are outside the listening area, especially at lateral positions very close to the loudspeaker, scenarios in which the listener can walk around the loudspeaker reveal effects of the directivity much more strongly and thus also allow a better evaluation of including source directivity, for example, concerning plausibility. Besides, further research is required to examine how accurately the directivity needs to be reproduced and how precisely it needs to be integrated into the synthesis (i.e., whether for example the early reflections also need to be adjusted according to the source directivity).

The physical evaluation presented in SEC. 2 revealed a good performance of the BRIR synthesis. At most of the tested positions in the room differences in DRR, $RT_{30}$, and IACC to reference measurements were below the respective JND, indicating inaudible differences, at least concerning the examined parameters. We obtained similarly good results with the three other sources in the same room and also tested the encoding with various RIRs of different rooms. However presenting all these results would go beyond the scope of this paper, so we decided to show a detailed physical evaluation based on the room in which we set up the plausibility experiment.

Overall the results suggest that the encoding, synthesis, and extrapolation work correctly. However the presented extrapolation treats early reflections as image sources and adjusts amplitude, TOAs, and DOAs accordingly when the listener moves through the room. Strictly speaking this is only correct if a) the selected early reflections originate from the surfaces of a shoebox room and not, for example, from reflecting objects in the room and b) the DOAs were correctly assigned to the selected reflections in the first place. Furthermore such an extrapolation of parameters is not necessarily correct in rooms with complex geometries. Therefore more research in different, more complex room geometries is required to further evaluate the extrapolation method and optimize the procedure for more general applicability.

In AR applications however the direction and distance of the direct sound are either predefined or can be estimated by using multimodal information, e.g., RGB or depth cameras, Simultaneous Localization and Mapping (SLAM), and can thus always be correctly adjusted according to the listener's movements and head orientation. Thus an incorrect extrapolation model would only negatively affect the early reflections. Additionally multimodal information could also be used to obtain rough estimations of room geometry and dimensions to inform the DOA generation process. However how precisely the early reflection pattern has to be adjusted according to the listener's movements to achieve a plausible reproduction and whether it always has to be dynamically adjusted at all needs further research.

The real-time framework presented in SEC. 3 allows experiencing a VAE with 6 DoF. The listener can move through the room, switching between real and virtual sources or changing the test signal, which allows an intuitive evaluation of the VAE. The system is not tied to the use of the *Paraspax* BRIR synthesis, i.e., any BRIRs can be used, and it can be flexibly extended, e.g., to implement real-time parametric rendering instead of convolution-based synthesis.

Based on the real-time framework we implemented a plausibility experiment according to Lindau and Weinzierl [40], which can be regarded as the strictest form of plausibility testing, as it is based on a comparison with real sound sources. According to this strict interpretation the 6-DoF VAE was plausible for one of the four expert listeners participating in the pilot study. Another participant had a detection rate only slightly above the critical value, whereas the other two participants could distinguish more reliably between virtual and real sources.

However due to the small number of participants and the fact that they were expert listeners we cannot yet draw any final conclusions, and we plan in future work to evaluate the perceptual plausibility of the presented 6-DoF VAE with a higher number of naive listeners. In particular we plan to examine the plausibility of the VAE for 6-DoF scenarios with different grid resolutions and trajectories, but we also plan to conduct 3-DoF experiments in which participants stand at different positions in the room and evaluate the plausibility of the system. Comparing those conditions allows us to investigate, for example, to what extent movement in a VAE can affect the perceived plausibility.

Interestingly the interviews with the participants of the pilot study revealed that often technical artifacts of the system made it possible to distinguish between a real and virtual source and not necessarily perceptible differences related to the synthesized BRIRs. Thus the pilot study showed that in order to achieve this strict form of plausibility, in which participants always compare with real sources, not only a high-quality BRIR synthesis is required but also a flawless technical framework, ideally without any artifacts that can be used by the listener to detect a virtual source. However, despite the artifacts described, the auralization is still very convincing and we therefore hypothesize that after solving the discussed technical challenges and further optimizing the real-time system, a plausible 6-DoF VAE based on one monaural RIR, tested as strictly as described, is feasible. To get an impression of the quality of the BRIR synthesis we provide some audio examples online.[2]

## 5 CONCLUSION

In this paper we presented a method for 6-DoF parametric spatial audio reproduction based on one monaural RIR. The *Paraspax* toolbox provides the entire pipeline to derive a spatial parametric description of the sound field from a monaural RIR and generate synthetic BRIRs for any desired head orientation and position in the room. The physical evaluation showed a good performance of the method, mostly with differences to reference measurements below the JND of the considered room acoustic parameter at all tested positions in the room. As a basis for a 6-DoF VAE we implemented a real-time framework that uses the synthesized BRIRs and enables both demo applications and psychoacoustic experiments in 6-DoF environments.

Using the framework we carried out a pilot study with expert listeners to assess the plausibility of the 6-DoF VAE. The results showed that the 6-DoF VAE in its current state can provide a plausible binaural reproduction for a listener moving through the room. However the results also revealed specific technical challenges for 6-DoF systems producing artifacts that are not directly related to the quality of the BRIR synthesis or the auralization but make it easier for participants to distinguish between real and virtual sources. Therefore our future work will mainly focus on optimizing the real-time framework to further enhance the plausibility of the 6-DoF VAE, as well as conducting listening experiments with naive listeners.

## 6 REFERENCES

[1] J. Merimaa and V. Pulkki, “Spatial Impulse Response Rendering I: Analysis and Synthesis,” *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127 (2005 Dec.).

[2] V. Pulkki, “Spatial Sound Reproduction With Directional Audio Coding,” *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516 (2007 Jun.).

[3] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, “Spatial Decomposition Method for Room Impulse Responses,” *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 17–28 (2013 Jan.).

[4] P. Stade, J. M. Arend, and C. Pörschmann, “Perceptual Evaluation of Synthetic Early Binaural Room Impulse Responses Based on a Parametric Model,” presented at the *142nd Convention of the Audio Engineering Society*, pp. 1–10 (2017 May), paper 9688.

[5] P. Coleman, A. Franck, D. Menzies, and P. J. B. Jackson, “Object-Based Reverberation Encoding From First-Order Ambisonic RIRs,” presented at the *142nd Convention of the Audio Engineering Society*, pp. 1–10 (2017 May), paper 9731.

[6] M. Zaunschirm, M. Frank, and F. Zotter, “BRIR Synthesis Using First-Order Microphone Arrays,” presented at the *144th Convention of the Audio Engineering Society*, pp. 1–10 (2018 May), paper 9944.

[7] T. Pihlajamäki and V. Pulkki, “Synthesis of Complex Sound Scenes With Transformation of Recorded Spatial Sound in Virtual Reality,” *J. Audio Eng. Soc.*, vol. 63, no. 7/8, pp. 542–551 (2015 Jul.). http://dx.doi.org/10.17743/jaes.2015.0059.

[8] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. A. P. Habets, “Six-Degrees-of-Freedom Binaural Audio Reproduction of First-Order Ambisonics With Distance Information,” in *Proceedings of the AES International Conference on Audio for Virtual and Augmented Reality*, pp. 1–10 (2018 Aug.), paper P6-2.

[9] J. G. Tylka and E. Y. Choueiri, “Domains of Practical Applicability for Parametric Interpolation Methods for Virtual Sound Field Navigation,” *J. Audio Eng. Soc.*, vol. 67, no. 11, pp. 882–893 (2019 Nov.). https://doi.org/10.17743/jaes.2019.0038.

[10] K. Müller and F. Zotter, “Auralization Based on Multi-Perspective Ambisonic Room Impulse Responses,” *Acta Acust.*, vol. 4, no. 6, pp. 1–18 (2020 Nov.). https://doi.org/10.1051/aacus/2020024.

[11] C. Pörschmann, P. Stade, and J. M. Arend, “Binauralization of Omnidirectional Room Impulse Responses - Algorithm and Technical Evaluation,” in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*, pp. 345–352 (Edinburgh, UK) (2017 Sep.).

[12] F. Brinkmann, H. Gamper, N. Raghuvanshi, and I. Tashev, “Towards Encoding Perceptually Salient Early Reflections for Parametric Spatial Audio Rendering,” presented at the *148th Convention of the Audio Engineering Society*, pp. 1–11 (2020 May), paper 10380.

[13] J. S. Abel and P. Huang, “A Simple, Robust Measure of Reverberation Echo Density,” presented at the *121st Convention of the Audio Engineering Society*, pp. 1–10 (2006 Oct.), paper 6985.

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

[15] G. Defrance, L. Daudet, and J.-D. Polack, “Finding the Onset of a Room Impulse Response: Straightforward?” *J. Acoust. Soc. Am.*, vol. 124, no. 4, pp. EL248–EL254 (2008 Oct.). https://doi.org/10.1121/1.2960935.

[16] S. Bech, “Timbral Aspects of Reproduced Sound in Small Rooms. I,” *J. Acoust. Soc. Am.*, vol. 97, no. 3, pp. 1717–1726 (1995 Mar.).

[17] B. Gourévitch and R. Brette, “The Impact of Early Reflections on Binaural Cues,” *J. Acoust. Soc. Am.*, vol. 132, no. 1, pp. 9–27 (2012 Jul.). https://doi.org/10.1121/1.4726052.

[18] S. E. Olive and F. E. Toole, “The Detection of Reflections in Typical Rooms,” presented at the *85th Convention of the Audio Engineering Society*, pp. 1–36 (1988 Nov.), paper 2719.

[19] P. Coleman, A. Franck, P. J. B. Jackson, R. J. Hughes, L. Remaggi, and F. Melchior, “Object-Based Reverberation for Spatial Audio,” *J. Audio Eng. Soc.*, vol. 65,

[2] Available: https://github.com/facebookresearch/Paraspax/tree/main/Paraspax_AudioExamples

no. 1/2, pp. 66–77 (2017 Jan.). https://doi.org/10.17743/jaes.2016.0059.

[20] E. Brandão, A. Lenzi, and S. Paul, "A Review of the *In Situ* Impedance and Sound Absorption Measurement Techniques," *Acta Acust. United Acust.*, vol. 101, no. 3, pp. 443–463 (2015 May/Jun.). https://doi.org/10.3813/AAA.918840.

[21] B. Friedlander and B. Porat, "The Modified Yule-Walker Method of ARMA Spectral Estimation," *IEEE Trans. Aero. Electr. Syst.*, vol. AES-20, no. 2, pp. 158–173 (1984 Mar.). https://doi.org/10.1109/TAES.1984.310437.

[22] J. M. Arend, T. Lübeck, and C. Pörschmann, "A Reactive Virtual Acoustic Environment for Interactive Immersive Audio," in *Proceedings of the AES International Conference on Immersive and Interactive Audio*, pp. 1–10 (2019 Mar.), paper 9.

[23] J. Ahrens, "Auralization of Omnidirectional Room Impulse Responses Based on the Spatial Decomposition Method and Synthetic Spatial Data," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 146–150 (Brighton, UK) (2019 May).

[24] S. Tervo, J. Pätynen, N. Kaplanis, M. Lydolf, S. Bech, and T. Lokki, "Spatial Analysis and Synthesis of Car Audio System and Car Cabin Acoustics With a Compact Microphone Array," *J. Audio Eng. Soc.*, vol. 63, no. 11, pp. 914–925 (2015 Nov.). https://doi.org/10.17743/jaes.2015.0080.

[25] L. McCormack, V. Pulkki, A. Politis, O. Scheuregger, and M. Marschall, "Higher-Order Spatial Impulse Response Rendering: Investigating the Perceived Effects of Spherical Order, Dedicated Diffuse Rendering, and Frequency Resolution," *J. Audio Eng. Soc.*, vol. 68, no. 5, pp. 338–354 (2020 May). https://doi.org/10.17743/jaes.2020.0026.

[26] S. V. Amengual Garí, J. M. Arend, P. T. Calamia, and P. W. Robinson, "Optimizations of the Spatial Decomposition Method for Binaural Reproduction," *J. Audio Eng. Soc.*, vol. 68, no. 12, pp. 959–976 (2020 Dec.). https://doi.org/10.17743/jaes.2020.0063.

[27] A. J. Kolarik, B. C. J. Moore, P. Zahorik, S. Cirstea, and S. Pardhan, "Auditory Distance Perception in Humans: A Review of Cues, Development, Neuronal Bases, and Effects of Sensory Loss," *Atten. Percept. Psych.*, vol. 78, no. 2, pp. 373–395 (2016 Feb.). https://doi.org/10.3758/s13414-015-1015-1.

[28] F. Menzer and C. Faller, "Investigations on an Early-Reflection-Free Model for BRIRs," *J. Audio Eng. Soc.*, vol. 58, no. 9, pp. 709–723 (2010 Sep.).

[29] T. Wendt, S. van de Par, and S. D. Ewert, "A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation," *J. Audio Eng. Soc.*, vol. 62, no. 11, pp. 748–766 (2014 Nov.). https://doi.org/10.17743/jaes.2014.0042.

[30] G. S. Kendall, "The Decorrelation of Audio Signals and Its Impact on Spatial Imagery," *Comp. Music J.*, vol. 19, no. 4, pp. 71–87 (1995). https://doi.org/10.2307/3680992.

[31] J. Blauert, *The Technology of Binaural Listening* (Springer-Verlag Berlin Heidelberg, Heidelberg, Germany, 2013).

[32] C. Borß and R. Martin, "An Improved Parametric Model for Perception-Based Design of Virtual Acoustics," in *Proceedings of the 35th International Conference: Audio for Games*, pp. 1–8 (2009 Feb.), paper 3.

[33] H. Kuttruff, *Room Acoustics (Fifth Edition)* (CRC Press, Boca Raton, FL, 2009).

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

[35] B. Xie, "On the Low Frequency Characteristics of Head-Related Transfer Function," *Chinese J. Acoust.*, vol. 28, no. 2, pp. 116–128 (2009).

[36] L. Aspöck, F. Brinkmann, D. Ackermann, S. Weinzierl, and M. Vorländer, "BRAS - Benchmark for Room Acoustical Simulation," http://dx.doi.org/10.14279/depositonce-6726.2 (2019).

[37] S. Csadi, F. M. Boland, L. Ferguson, H. O'Dwyer, and E. Bates, "Direct to Reverberant Ratio Measurements in Small and Mid-Sized Rooms," in *Proceedings of the AES International Conference on Immersive and Interactive Audio*, pp. 1–10 (2019 Mar.), paper 28.

[38] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the Minimum Audible Difference in Direct-to-Reverberant Energy Ratio," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 450–461 (2008 Jul.). https://doi.org/10.1121/1.2936368.

[39] H. Seraphim, "Untersuchungen über die Unterschiedsschwelle exponentiellen Abklingens von Rauschbandimpulsen," *Acta Acust. United Acust.*, vol. 8, supp. 1, pp. 280–284 (1958).

[40] A. Lindau and S. Weinzierl, "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acust. United Acust.*, vol. 98, no. 5, pp. 804–810 (2012 Sep./Oct.). https://doi.org/10.3813/AAA.918562.

[41] A. Neidhardt, F. Klein, N. Knoop, and T. Köllmer, "Flexible Python Tool for Dynamic Binaural Synthesis Applications," presented at the *142nd Convention of the Audio Engineering Society*, pp. 1–5 (2017 May), paper 346.

[42] J. Gómez-Bolaños, A. Mäkivirta, and V. Pulkki, "Automatic Regularization Parameter for Headphone Transfer Function Inversion," *J. Audio Eng. Soc.*, vol. 64, no. 10, pp. 752–761 (2016 Oct.). http://dx.doi.org/10.17743/jaes.2016.0030.

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

[44] T. Waltemate, F. Hülsmann, T. Pfeiffer, S. Kopp, and M. Botsch, "Realizing a Low-Latency Virtual Reality Environment for Motor Learning," in *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology*, pp. 139–147 (2015 Nov.). https://doi.org/10.1145/2821592.2821607.

[45] T. Robotham, O. Rummukainen, J. Herre, and E. A. P. Habets, "Evaluation of Binaural Renderers in Virtual

Reality Environments: Platform and Examples," presented at the *145th Convention of the Audio Engineering Society*, pp. 1–5 (2018 Oct.), paper 454.

[46] S. V. Amengual Garí, W. O. Brimijoin, H. G. Hassager, and P. W. Robinson, "Flexible Binaural Resynthesis of Room Impulse Responses for Augmented Reality Research," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, pp. 161–166 (Paris, France) (2019 Sep.). https://doi.org/10.25836/sasp.2019.31.

[47] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A Round Robin on Room Acoustical Simulation and Auralization," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. 2746–2760 (2019 Apr.). https://doi.org/10.1121/1.5096178.

[48] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the Authenticity of Individual Dynamic Binaural Synthesis," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 1784–1795 (2017 Oct.). https://doi.org/10.1121/1.5005606.

[49] S. Werner, F. Klein, and G. Götz, "Investigation on Spatial Auditory Perception using Non-Uniform Spatial Distribution of Binaural Room Impulse Responses," in *Proceedings of the 5th International Conference on Spatial Audio (ICSA)*, pp. 137–144 (2019 Sep.). https://doi.org/10.22032/dbt.39967.

[50] C. Schissler, P. Stirling, and R. Mehra, "Efficient Construction of the Spatial Room Impulse Response," in *Proceedings of the IEEE Virtual Reality (VR)*, pp. 122–130 (Los Angeles, CA) (2017 Mar.). https://doi.org/10.1109/VR.2017.7892239.

[51] U. Sloma, F. Klein, S. Werner, and T. Pappachan Kannookadan, "Synthesis of Binaural Room Impulse Responses for Different Listening Positions Considering the Source Directivity," presented at the *147th Convention of the Audio Engineering Society*, pp. 1–9 (2019 Oct.), paper 10237.

[52] T. Robotham, O. S. Rummukainen, and E. A. P. Habets, "Towards the Perception of Sound Source Directivity Inside Six-Degrees-of-Freedom Virtual Reality," in *Proceedings of the 5th International Conference on Spatial Audio (ICSA)*, pp. 71–78 (Ilmenau, Germany) (2019 Sep.). https://doi.org/10.22032/dbt.39956.

## APPENDIX

Table 1. Estimated DRR, $RT_{30}$, and IACC values for the measured RIRs ($RIR_M$), measured BRIRs ($BRIR_M$), and synthesized BRIRs ($BRIR_S$) at the grid positions.

| Position | DRR in dB | | | $RT_{30}$ in s | | | IACC | |
|---|---|---|---|---|---|---|---|---|
| | $RIR_M$ | $BRIR_M$ | $BRIR_S$ | $RIR_M$ | $BRIR_M$ | $BRIR_S$ | $BRIR_M$ | $BRIR_S$ |
| 1 | −4.73 | −3.10 | −3.73 | 0.90 | 0.85 | 0.92 | 0.21 | 0.14 |
| 2 | −2.59 | ··· | −2.89 | 0.88 | ··· | 0.91 | ··· | 0.38 |
| 3 | −2.62 | −1.86 | −2.58 | 0.90 | 0.84 | 0.91 | 0.40 | 0.43 |
| 4 | −4.48 | ··· | −2.97 | 0.91 | ··· | 0.91 | ··· | 0.25 |
| 5 | −6.56 | ··· | −5.08 | 0.88 | ··· | 0.92 | ··· | 0.19 |
| 6 | −5.18 | −3.72 | −4.61 | 0.93 | 0.84 | 0.92 | 0.31 | 0.23 |
| 7 | −4.83 | −4.16 | −4.33 | 0.90 | 0.83 | 0.92 | 0.34 | 0.36 |
| 8 | −6.10 | ··· | −4.42 | 0.91 | ··· | 0.92 | ··· | 0.24 |
| 9 | −8.27 | −6.47 | −6.48 | 0.88 | 0.82 | 0.91 | 0.18 | 0.18 |
| 10 | −7.14 | −5.89 | −6.05 | 0.90 | 0.84 | 0.92 | 0.29 | 0.35 |
| 11 | −7.04 | −6.00 | −5.89 | 0.90 | 0.82 | 0.92 | 0.26 | 0.35 |
| 12 | −7.61 | −6.61 | −5.92 | 0.90 | 0.84 | 0.91 | 0.16 | 0.24 |
| 13 | −9.49 | ··· | −7.25 | 0.89 | ··· | 0.92 | ··· | 0.17 |
| 14 | −8.78 | ··· | −7.21 | 0.91 | ··· | 0.91 | ··· | 0.22 |
| 15 | −8.90 | −7.76 | −7.02 | 0.92 | 0.83 | 0.92 | 0.24 | 0.32 |
| 16 | −9.20 | −7.95 | −6.87 | 0.89 | 0.83 | 0.92 | 0.17 | 0.24 |
| 17 | −10.74 | ··· | −8.15 | 0.87 | ··· | 0.92 | ··· | 0.20 |
| 18 | −9.84 | −8.26 | −8.13 | 0.92 | 0.85 | 0.92 | 0.26 | 0.32 |
| 19 | −10.09 | −9.06 | −8.18 | 0.89 | 0.83 | 0.92 | 0.18 | 0.32 |
| 20 | −10.25 | ··· | −7.73 | 0.89 | ··· | 0.92 | ··· | 0.23 |

DRR calculated with a direct sound window of 1.5 ms. For BRIRs, the direct and diffuse sound energy of both channels were summed before DRR estimation. $RT_{30}$ calculated as the mean in the octave bands from 0.25–8 kHz. For BRIRs, determined broadband $RT_{30}$ values per channel were averaged to obtain a single value. Missing values due to missing measurement data indicated by (···).

## THE AUTHORS

Johannes M. Arend

Sebastià V. Amengual Garí

Carl Schissler

Florian Klein

Philip W. Robinson

Johannes M. Arend received a B.Eng. degree in media technology from HS Düsseldorf (Germany) in 2011 and an M.Sc. degree in media technology from TH Köln (Germany) in 2014. Since 2015 he has been a Research Fellow and working toward a Ph.D. degree at TH Köln and TU Berlin in the field of spatial audio with a focus on binaural technology, auditory perception, and audio signal processing. Between September 2019 and March 2020 he was a research intern at Facebook Reality Labs Research.

•

Sebastià V. Amengual Garí is currently a research scientist at Facebook Reality Labs Research working on room acoustics, spatial audio, and auditory perception. He received a Diploma Degree in Telecommunications with a major in Sound and Image in 2014 from the Polytechnic University of Catalonia (UPC) in 2014, completing his Master's Thesis at the Norwegian University of Science and Technology (NTNU). His doctoral work at the Detmold University of Music focused on investigating the interaction of room acoustics and live music performance using virtual acoustic environments. His research interests lie in the intersection of audio, perception, and music.

•

Carl Schissler is currently a research scientist at Facebook Reality Labs Research, which he joined in 2017 after receiving his Ph.D. in computer science from the University of North Carolina at Chapel Hill. His primary research interests include real-time acoustic simulation, digital signal processing, and computational geometry. Beyond research, Carl is a multi-instrumentalist and amateur composer and enjoys audio mixing for recording and stage.

•

Florian Klein is a Ph.D. student at Technical University Ilmenau, Germany. His main research area is auditory adaptation processes in the scope of spatial hearing. Furthermore he is working on solutions for 6DOF spatial sound rendering for AR/VR applications. He received the Best Student Paper Award at the AES Convention 2015 and a Best Paper Award at the AES Convention 2019. Between November 2019 and May 2020 he was a research intern at Facebook Reality Labs Research.

•

Philip W. Robinson is a research science manager in room acoustics and auditory perception at Facebook Reality Labs Research (FRL Research) in Redmond, WA. Prior to joining FRL Research he incorporated virtual acoustics simulation and reproduction systems into building design processes at the architecture firm of Foster + Partners. He was a Fulbright Scholar and post-doctoral researcher at Aalto University in Finland, where he studied perception of concert hall acoustics, spatial auditory resolution, and echo thresholds. He has been a visiting researcher at EPFL in Switzerland and Hanyang University in South Korea. He received a Ph.D. from Rensselaer Polytechnic Institute in Troy, NY in 2012. In a previous life he was a registered architect in his home state of New Mexico. He remains passionate about architecture, the study of which gave him a great interest in perception of environments, real or virtual.