T. Jüterbock, F. Brinkmann, H. Gamper, N. Raghuvanshi
and S. Weinzierl, "Spatio-Temporal Windowing for Encoding
Perceptually Salient Early Reflections in
Parametric Spatial Audio Rendering"
J. Audio Eng. Soc., vol. 71, no. 10, pp. 664–678, (2023 October).
DOI: https://doi.org/10.17743/jaes.2022.0099

Spatio-Temporal Windowing for Encoding Perceptually Salient Early Reflections in Parametric Spatial Audio Rendering

TOBIAS JÜTERBOCK,[1] **FABIAN BRINKMANN,**[1] *AES Associate Member*,
(a.jueterbock@tu-berlin.de) (fabian.brinkmann@tu-berlin.com)
HANNES GAMPER,[2] *AES Member*, **NIKUNJ RAGHUVANSHI,**[2] *AES Associate Member* **AND**
(hannes.gamper@microsoft.com) (nikunjr@microsoft.com)
STEFAN WEINZIERL[1]
(stefan.weinzierl@tu-berlin.de)

[1]*Audio Communication Group, Technical University of Berlin, Berlin, Germany*
[2]*Microsoft Research Redmond, Redmond, WA*

Parametric spatial audio rendering aims to provide perceptually convincing audio cues that are agnostic to the playback system to enable the acoustic design of games and virtual reality. The authors propose an algorithm for detecting perceptually important reflections from spatial room impulse responses. First, a parametric representation of the sound field is derived based on perceptually motivated spatio-temporal windowing, followed by a second step that estimates the perceptual salience of the detected reflections by means of a masking threshold. In this work, a vertical dependency is incorporated into both these components. This was inspired by recent research revealing that two sound sources in the median plane can evoke two independent auditory events if their spatial separation is sufficiently large. The proposed algorithm is evaluated in nine simulated shoebox rooms with a wide range of sizes and reverberation times. Evaluation results show improved selection of early reflections by accounting for source elevation and suggest that for speech signals, the perceptual quality increases with an increasing number of rendered early reflections.

0 INTRODUCTION

Virtual reality (VR), augmented reality (AR), and gaming applications must perform 3D sound rendering within a small fraction of a single CPU core because resources are typically shared with other compute-intensive aspects of a full system, including visual rendering and character animation. At the same time, the audio rendering must remain perceptually plausible to provide consistent audio-visual cues that can enhance the sense of presence and immersion. One approach to meet these opposing goals is a parametric representation of spatial sound fields that estimates perceptually relevant aspects in an offline encoding step and efficiently decodes to 3D sound in real time.

Common parametric models include various aspects such as the time of arrival (TOA), amplitude, and direction of arrival (DOA) of the first sound and early reflections, as well as a description of the late reverberation in terms of its level and decay [1–5]. These models require algorithms to automatically extract a small set of perceptually salient early reflections given a spatial room impulse response (SRIR) for fast rendering.

Coleman et al. encoded the six loudest reflections detected from SRIRs captured with 48 microphones [1]. They used the Clustered Dynamic Programming Projected Phase-Slope Algorithm [6] to extract the TOA of the six strongest peaks from the multichannel RIRs and then applied delay-and-sum beamformers to a time window of 1.3 ms around the TOA to estimate the DOA and level of each reflection. When using first-order Ambisonics RIRs, they detected the 20 loudest peaks [2], applying the mono-channel Dynamic Programming Projected Phase-Slope Algorithm [7] to the omnidirectional channel to extract the TOAs. A time window of 1.3 ms was applied around each TOA, and a virtual cardioid microphone was steered toward the maximum energy of each window to estimate the DOAs and levels of the reflections.

Stade et al. [3] used between 50 and 200 reflections encoded from 1,202 microphones to synthesize binaural RIRs. They detected reflections by analyzing intensity matrices obtained from a plane wave decomposition of the sound field. The intensity matrices were calculated for different points in time based on short-time Fourier transforms. Reflections were selected by identifying local maxima in the

intensity matrix, comparing them to their spatial neighborhood to cluster or separate maxima to single or multiple reflections, and finally comparing their amplitudes to an absolute threshold derived from the Energy Decay Curve.

In these studies, the parametric renderings were of high quality but still discernible from the reference, partly because the perceptual tests did not isolate the effect of rendering the early reflections. Müller and Zotter [8] proposed a six-degrees-of-freedom Ambisonic RIRs rendering plugin, detecting peaks in the Ambisonic RIR by running a 0.5 ms Hamming-windowed moving-average filter across the amplitude of the pseudo intensity vector and extracting directional sound events by applying asymmetrical cosine-shaped window functions (spanning from 0.5 ms before the TOA to the next detected peak, up to 5 ms after the TOA) to the directional component of each segment. In their work, the ten loudest reflections are considered for further processing.

As can be seen from the above, reflections are often detected solely based on temporal properties of single-channel impulse responses without considering the spatial resolution of the human auditory system. Moreover, the number of rendered reflections is often fixed without considering their perceptual relevance. Few studies determined and, in some cases, discarded inaudible reflections based on listening tests [9–11] or binaural models [12, 13]. A drawback of these methods is that they assume the TOA, amplitude, and DOA of the first sound and reflections to be known and that they use computationally expensive processing, including convolution and filter banks. Röhrbein and Lindau [13] found that for speech and music content, 5–11 reflections may be sufficient for perceptually transparent rendering. For highly artificial Dirac pulse trains, rendering 76 reflections was, however, still distinguishable from the reference. It might thus be expected that transient-rich natural sounds such as castanets, clapping, or gun sounds require rendering more than 5–11 reflections, a content dependency that is better documented for the simple case of a direct sound followed by a single reflection [14, 12].

The authors propose a two-step algorithm that aims at combining the two approaches discussed above. In the first step, a parametric representation of the sound field is derived based on perceptually motivated spatio-temporal windowing, followed by a second step that estimates the perceptual salience of the detected reflections by means of a masking threshold. This is used to reduce the number of reflections being rendered at a later stage, without necessarily being limited to a fixed number of reflections. The approach is evaluated with respect to potential audible differences caused by the masking threshold and the effect of further reducing the amount of rendered reflections. A possible use case is the automatic offline encoding of billions of spatially distributed listener and source position pairs [4].

This publication is based on previous work that introduced the general encoding and decoding framework [15].[1] In the current study, this framework is extended to account for the effect of source elevation in the spatio-temporal windowing and masking threshold.

1 BACKGROUND

The German Standard DIN 1320 defines audible sound as "mechanical vibrations and waves of an elastic medium in the frequency range of human hearing" [16, 17]. Following this definition, Blauert [17] defines a *sound event* as a physical event involving sound, e.g., the sound waves emitted by a speaker.

In contrast, what is perceived by humans is defined by Blauert as an *auditory event*. Auditory events are often caused, determined, or elicited by sound events. However, auditory events can occur without a sound event (e.g., tinnitus), sound events do not necessarily result in auditory events (e.g., sound events below the hearing threshold), and multiple sound events might be perceived as one auditory event. In the context of room acoustics, the first sound, early and late reflections can be described as sound events with a TOA and DOA and a (frequency-dependent) amplitude and phase response. Corresponding auditory events may have other psychoacoustic qualities like loudness (not necessarily equal but correlated to the amplitude), perceived width, pitch, or distance. For human listeners, it is fair to assume that the number of auditory events necessary to capture the room acoustic impression is significantly lower than the number of sound events necessary to fully reproduce the physical sound field.

To obtain a better understanding of auditory events in the context of room acoustics, the available literature is briefly reviewed in the following. An interaural polar coordinate system shown in Fig. 1 was used because it corresponds with the mechanism of the auditory system that uses binaural cues—the interaural time and level differences—for localization in the lateral dimension and monaural spectral cues for localization in the polar dimension.

1.1 Spatio-Temporal Window

Temporal aspects of early reflections are relatively well studied as part of the precedence effect and summing localization [17–19]. In this context, it was shown that the auditory system averages incoming sound into a single auditory event up to about 1 ms after the first sound [17, chapter 3.1]. However, comb filter effects become particularly noticeable for reflections with a delay of 0.5–2 ms at least for some signal types [20].

Fewer empirical data are available for the spatial aspects of early reflections. Best et al. investigated the ability of human listeners to perceive multiple simultaneous, equally strong broadband sound sources in the frontal horizontal

[1] This study claimed that rendering the six first-order reflections was indistinguishable from the reference. The authors would like to clarify that this was not the case but a mistake in the preparation of the stimuli for the experiment. Informal listening after noting the error suggested that rendering the six first-order reflections was comparable to rendering the six loudest reflections detected by the masking threshold.

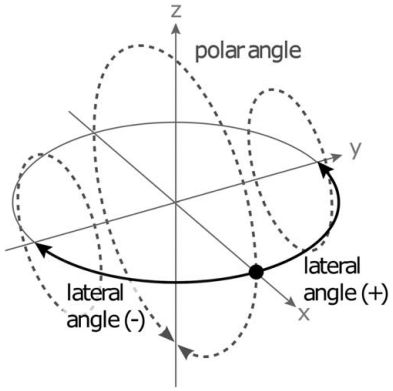

Fig. 1. Interaural polar coordinate system with the lateral angle φ and polar angle θ. For $\varphi = 0$, $\theta = \{0, 90^\circ, 180^\circ, -90^\circ\}$ denotes sources in front, above, behind, and below the listener. The dashed circles show directions with constant lateral angle, which are each located on sagittal planes. Note that $|\varphi| \leq 90^\circ$, and for sources from the left and right side ($\varphi = \pm 90^\circ$), all polar angles collapse to one point.

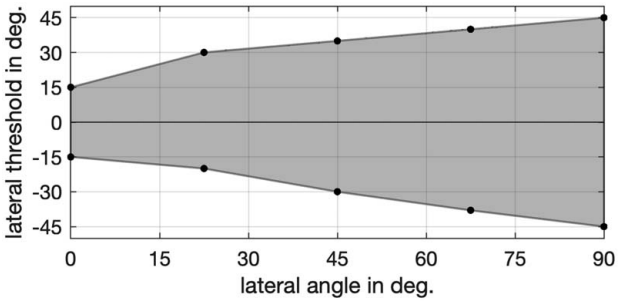

Fig. 2. Threshold for source separation in the horizontal plane as a function of the lateral angle of the target source. The values shown by the dots were extracted from [21, Fig. 3(e)] and linearly interpolated. The gray area indicates the spatial window outside of which two simultaneous equally strong sound sources can be perceived as individual auditory events.

plane as individual auditory events [21]. They found that the auditory system uses binaural cues to separate sources in the lateral direction and that the angular separation that is required for perceiving two sound sources increases with the lateral angle. The detected angular threshold for source separation is shown in Fig. 2.

Even fewer data are available for the perception of simultaneous sources in the polar dimension. For the frontal median sagittal plane, Bremen et al. and Pulkki et al. found bimodal localization responses for two concurrent sound sources at opening angles larger than 45°–60° [22, 23].

1.2 Spatio-Temporal Masking

The masking threshold (or reflection masked threshold) defines the threshold above which a human listener is able to perceive a room reflection by means of perceived changes in timbre, loudness, or spatial aspects. The temporal dependency of the threshold was extensively investigated as part of the precedence effect for the simple case of two sound events. In this case, the threshold exhibits an exponential

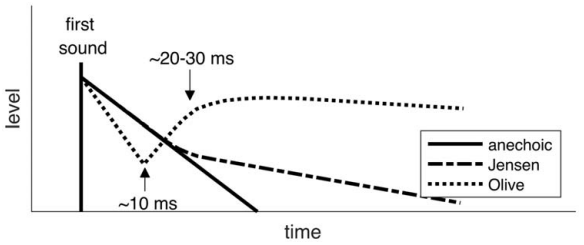

Fig. 3. Schematic overview of masking threshold in anechoic (solid line) and reverberant environments (dashed lines). See SEC. 1.2 for more information.

decay as a function of the delay of the second sound event [12, 24]. In reverberant environments, the reflections themselves act as maskers, generally decreasing the decay rate of the masking threshold over time [14, 11]. However, the exact shape of the threshold differs across studies. Whereas data from Olive and Toole [14] suggest a v-shaped threshold that exhibits an initial steep negative slope that transitions to a positive slope in the presence of reverberant energy, experiments from Jensen and Welti [11] show a simpler double-sloped threshold (cf. Fig. 3).

A spatial dependency of the masking threshold can be observed in previous studies [9, 10, 12, 25–27], albeit with some differences regarding the exact nature of the dependence. The studies show a decrease of the threshold in the range of 7–15 dB (depending on the audio content) for reflections that are spatially separated from the first sound.

2 EXTRAPOLATING THE POLAR THRESHOLD FOR SOURCE SEPARATION

A key component of this work is the use of spatio-temporal windowing to parameterize SRIRs into auditory events. This involves defining a spatial window in both lateral and polar directions within which no simultaneous, separate auditory events can be perceived, so that all samples lying within the window can be assigned to a common auditory event. Because auditory events arrive from all directions, this spatial window must be well-defined for all directions of incidence. As discussed in SEC. 1.1, the angular threshold for source separation in the lateral direction is mainly determined by binaural cues and is, therefore, approximately rotationally symmetric about the interaural axis. Hence, the window width in the lateral direction is described completely by Fig. 2.

In contrast, binaural cues cannot be exploited on sagittal planes (cones of confusion). Thus, it can be assumed that the auditory system uses monaural spectral cues to discriminate simultaneous sources in the polar dimension. These spectral cues change not only depending on the polar but also on the lateral angle of incidence, so that the window width in the polar direction cannot be assumed to be rotationally symmetric. Consequently, it is not trivial to extrapolate the available empirical data from the sagittal median plane to other directions of incidence.

The extrapolation method is divided into three steps: First, a metric is developed to measure the perceived difference between a single and two simultaneously active sources in the sagittal median plane (SEC. 2.1 and 2.2). Second, the conditions in which two sources are perceived as separate auditory events are reproduced, and a threshold is derived using the metric from the first step (SEC. 2.3). Third, the minimum spatial separation in which this threshold is exceeded is calculated (SEC. 2.3). These opening angles at the threshold can then be used as the spatial window widths in the polar direction.

This implies the assumptions that the perceived differences between the spectra increase monotonically beyond this point and that all concurrent sources with larger spatial separation can also be separated perceptually. This is not necessarily the case in general, but because all contributions from the spatial windows derived from this threshold would be combined into one auditory event, the priority was to conservatively determine the smallest opening angle at which two sources can be separated in order to prevent erroneous merging of auditory events.

2.1 Modeling the Perceived Location of Concurrent Sources

The probabilistic median plane localization model from Baumgartner et al. [28] was chosen as a metric to estimate the perceived difference between one and two active sound sources (the model is contained in the Auditory Modeling Toolbox [29]). The model compares a target head-related transfer function (HRTF) set to a set of template HRTFs and estimates a probability density function (PDF) for the perceived source elevation. The target HRTF set was obtained by summing two HRTFs: The first HRTF was kept fixed at the reference source position (φ_{ref}, θ_{ref}), while the position of the second HRTF was moved around a cone of confusion, i.e., keeping the lateral angle constant while varying the polar angle ($\varphi_{\text{test}} = \varphi_{\text{ref}}$, $\theta_{\text{test}} = \theta_{\text{ref}} + \Delta\theta$, where $\Delta\theta$ is the opening angle). This procedure was applied to an equidistant grid of reference positions φ_{ref} and θ_{ref} (in 5° steps) and opening angles $\Delta\theta$ (in 4° steps). The resulting PDFs were averaged across 94 unique subjects contained in the HUTUBS database [30].

Selected results for the incidence angle $\varphi_{\text{ref}} = 0$, $\theta_{\text{ref}} = -50°$ are shown in Fig. 4(a). For an opening angle of $\Delta\theta = 0$, reference and target HRTFs are identical. As expected, the resulting PDF, modeling the perceived position, has a narrow and distinct peak at the reference position. At opening angles of $\Delta\theta = 30°$ and $\Delta\theta = 60°$, the peaks become progressively wider and show their maximum between the reference position and $\Delta\theta$. In agreement with Bremen et al. [22], this suggests that the model predicts a perceived phantom sound source that is less accurately localized in these cases. At larger opening angles ($\Delta\theta = 90°$ in this example), the model no longer estimates a dominant perceived source location. For these cases, multi-modal distributions start to emerge, although they are not as clear as findings shown by Pulkki et al. [23].

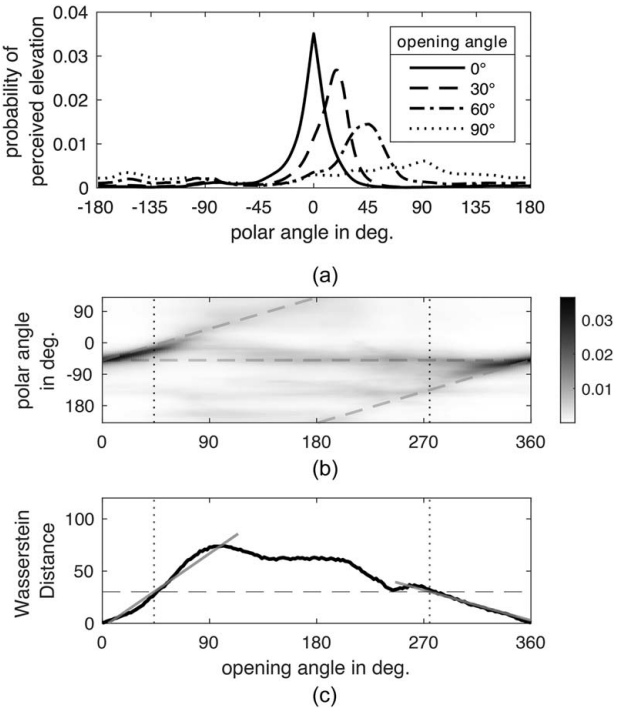

Fig. 4. Calculation of the polar threshold for source separation at $\varphi_{\text{ref}} = 0$, $\theta_{\text{ref}} = -50°$, averaged across 94 HRTF sets. (a) PDFs for the opening angles $\Delta\theta = \{0°, 30°, 60°, 90°\}$. (b) PDFs for all opening angles. Each vertical slice of the surface plot represents one probability distribution as shown above. The dashed lines indicate the two source locations. (c) The Wasserstein distance W_1 between all PDFs and the PDF for $\Delta\theta = 0°$ (single source). The solid gray line shows the linear fit of W_1, the dashed gray line shows the assumed W_1-threshold [cf. Eq. (1)], and the dotted gray lines show the estimated polar thresholds for source separation, obtained by intersecting linear fits with the W_1-threshold.

The PDFs for the same reference position but all opening angles $\Delta\theta$ are shown in Fig. 4(b). The PDFs of Fig. 4(a) can be found at opening angles (x-values) of 0, 30°, 60°, and 90°. As described above, the PDFs show clear peaks for small to medium opening angles indicating the perception of a single (phantom) source.

Two-dimensional PDFs, as shown in Fig. 4, were computed for all reference angles and were qualitatively comparable to the discussed example. The model's results are generally consistent with previous studies on the perception of multiple sources in the median plane. However, the model suggests that the spatial separation required for multi-modal distributions to become apparent may be larger than previously reported in studies such as [22, 23].

2.2 Evaluating the Difference Between Probability Distributions

Subsequently, the differences between the probability distributions for $\Delta\theta \neq 0$ and the single source case $\Delta\theta = 0$ had to be quantified. The first Wasserstein distance was chosen because it produced smooth curves that were consistent across opening angles and reference positions. It is a similarity measure for probability distributions that is also known as the Earth Mover's Distance. Interpreting two

probability distributions as piles of earth, the Wasserstein distance gives the minimum amount of earth that has to be moved in order to convert one probability distribution into the other. Directly calculating the Wasserstein distance using Pele's implementation [31] yielded almost identical results as generating random numbers from the distributions [32] and then calculating the Wasserstein distance using Kolbe's implementation [33]. The second approach was computationally more efficient and was therefore used.

The resulting one-dimensional function of the opening angle, shown in black in Fig. 4(c), was obtained through this process. For $\Delta\theta = 0 = 360^\circ$, the Wasserstein distance always tends to zero because the corresponding PDF is compared to itself. With increasing opening angles, the Wasserstein distance usually increased monotonically up to a certain point. To further smooth the curves, separate linear least-squares fits were performed for positive and negative opening angles, shown as gray lines in Fig. 4(c). The fitting ranges were set to the intervals $\Delta\theta = [0, \pm 114^\circ]$ in order to minimize the overall fitting error. Varying the intervals around the optimum value had no significant influence on the results.

2.3 Applying the Threshold

To relate the Wasserstein distance to the threshold for source separation, the configuration that evoked bimodal localization in Bremen et al. [22] was reproduced ($\varphi_{\text{ref}} = 0$, $\theta_{\text{ref}} = -22.5^\circ$, $\Delta\theta = 45^\circ$), resulting in a fitted Wasserstein Distance of $t_W = 30.2332$. This value was then assumed to be the threshold above which humans can perceive separate co-occurring sound sources in the sagittal median plane. In the lateral-polar coordinate system used, the arc length between adjacent polar coordinate points decreases by the factor $\cos(\varphi)$ for non-zero lateral angles, which affects the Wasserstein distance in a similar manner. To account for this, the threshold for the sagittal median plane was increased for non-zero lateral angles in a final step:

$$t(\varphi_{\text{ref}}) = \frac{t_W}{\cos(\varphi_{\text{ref}})}. \tag{1}$$

Based on this, the fitted Wasserstein distance curves [solid gray lines in Fig. 4(c)] were intersected with $t(\varphi_{\text{ref}})$ to determine the polar threshold for source separation—i.e., the minimum opening angle $\Delta\theta$ required to perceive two sound sources as separate auditory events—for all reference positions. Fig. 4(c) illustrates this process for one reference position.

Results for the polar threshold for source separation for all reference positions are shown in Fig. 5. The right and left-hand sides show the results for positive and negative opening angles, respectively. The results from Fig. 4(c) can be found here by examining the surface plot at the absolute lateral reference angle 0 and the polar reference angle -50°. The lowest polar threshold values appear in the median plane and increase with the absolute lateral angle. Different results are obtained depending on whether the opening angle is positive or negative due to the asymmetry of the underlying HRTF sets.

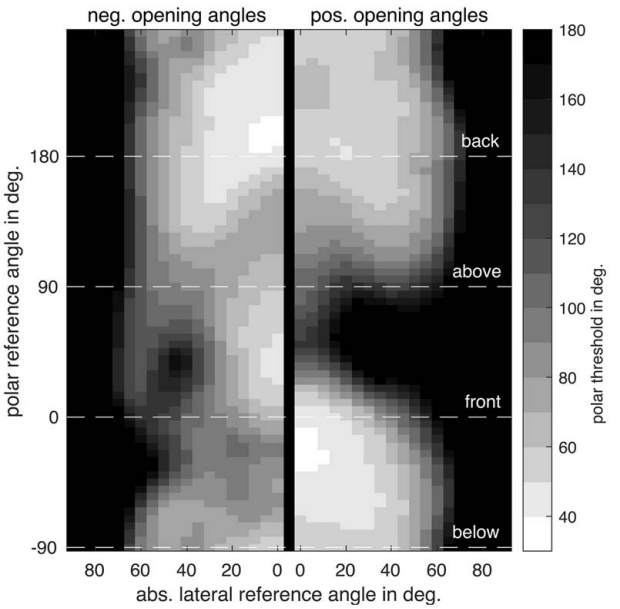

Fig. 5. Extrapolated values of the polar threshold for source separation (a) for negative opening angles and (b) for positive opening angles. Opening angles are wrapped to $(-180^\circ, 180^\circ]$.

Furthermore, the thresholding of the Wasserstein Distance curves takes into account not only the reference and test angles but also all directions in between, leading to direction-dependent results. For positive opening angles, the polar threshold is smaller for the frontal and rear median plane as compared to the upper median plane. For negative opening angles in the upward direction, the Wasserstein Distance often plateaued slightly below the threshold value, causing the linear fit to underestimate the resulting polar threshold in that region (except for the narrow peak seen on the left side in Fig. 5 at $|\varphi| = 45^\circ$, $\theta = 40^\circ$). This was not deemed to be problematic because it can only lead to more conservative detection of reflections, thus never wrongly discarding or grouping important reflections.

3 ENCODING AND DECODING

This section describes the proposed pipeline to arrive at a parametric SRIR representation. As depicted in Fig. 6, the first step of the pipeline is the segmentation of the SRIR—a waveform representing the sound events—into a list of parametric auditory events—the reflections—by applying perceptually motivated spatio-temporal search windows according to SEC. 1.1. These auditory events are considered to be audible when occurring in isolation. The audibility of a single event in the presence of the remaining events is assessed in the second step by means of a masking threshold considering the spatio-temporal aspects described in SEC. 1.2. To account for the remaining energy in the SRIR, a parametric late reverberation is generated in the last step, which could be interpreted as a non-directional auditory event.

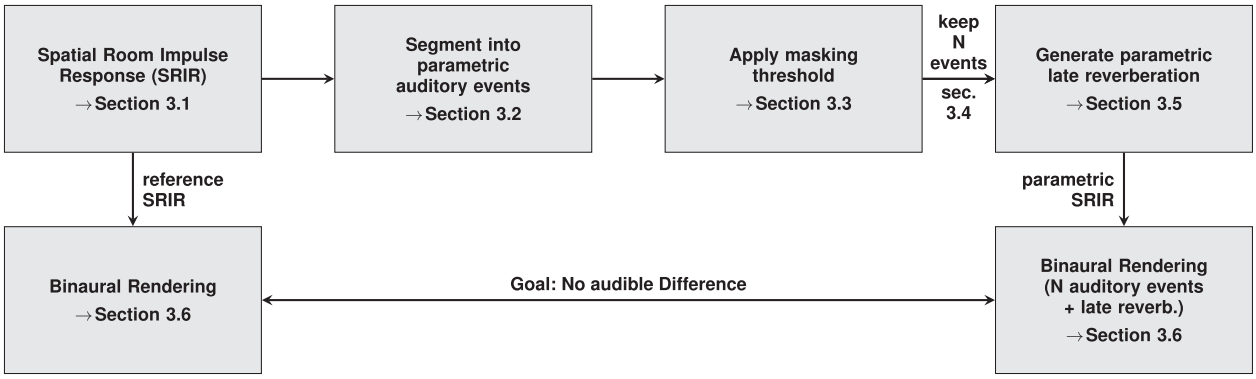

Fig. 6. Flowchart of encoding and decoding chain.

At the end of the pipeline, parametric SRIRs were evaluated against the reference by means of physical and perceptual analyses.[2]

3.1 Spatial Room Impulse Responses

A dataset of SRIRs was generated for testing. It was computed based on a hybrid room acoustical simulation using an image source model (ISM) for the early reflections and stochastic decaying noise for the late reverberation [34]. This dataset served as a reference for the perceptual evaluation to ensure that differences between the reference and the parametric approach can almost exclusively be attributed to differences in the rendering of early reflections. For the sake of simplicity, all simulations used frequency-independent boundary reflectivity and omnidirectional sources. The hybrid ISM model was used to generate SRIRs for nine shoebox-shaped empty rooms for all combinations of three room volumes $\mathrm{V} = \{200;\ 1{,}000;\ 5{,}000\}\ \mathrm{m}^3$ and reverberation times $T_{60} = \{0.5, 1, 2\}$ s. The ratio of each room's length, width, and height was set to 1.9:1.4:1. Uniform absorption coefficients were calculated according to Sabine's formula to match the target reverberation times.

To make sure that all perceptually relevant early reflections are included in the simulation, the image source model was used up to 1.5 times the estimated perceptual mixing time given by $\mathrm{T_{mix}} = 0.0117\,\mathrm{V} + 50.1$ ms [35]. The late reverberation was modeled as decaying white Gaussian noise, and its sample-wise DOA was drawn from a uniform random distribution. An exponential fade-in that started at the position of the direct sound with a level of –60 dB with respect to the level at $1.5\,\mathrm{T_{mix}}$ was applied to the late reverberation to achieve a smooth transition between the early and late part. Three receiver positions were considered per room (cf. Fig. 7): (i) at a distance of two times the critical distance [36, Eq. (5.39)] with respect to $T_{60} = 0.5$ s from the source close to the center of the room; (ii) 1 m from a wall, to get a strong back wall reflection; and (iii) 1 m from two walls in a corner, to get strong second-order reflections. Sources and receivers were positioned at a height of 1.6 m.

[2] Sample code, SRIRs, and binaural renderings are available at https://github.com/microsoft/Perceptual_saliency_of_early_reflections.

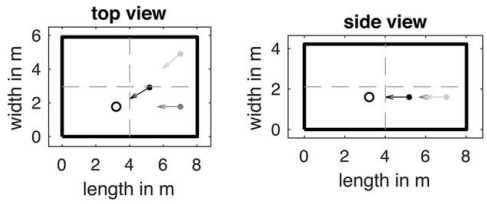

Fig. 7. Sketches of the small ISM room including the source position (circle), receiver positions (dots), receiver viewing directions (arrows), and symmetry axes (dashed).

3.2 SRIR Segmentation

First, the early part of the SRIR between the direct sound and $t \leq 200$ ms after the direct sound was segmented into auditory events. The direct sound TOA was estimated using a first-moment onset detector based on the cumulative energy of the omnidirectional SRIR channel p_{omni}, which proved to provide spatially smooth estimates for large numbers of source/receiver positions [4, Eq. (15)]. The 200 ms were used as a conservative value that well exceeds estimates of the perceptual mixing time [35, 37], which could alternatively be used as an upper limit.

The segmentation was done iteratively by (1) finding the sample in the SRIR with the largest absolute value; (2) assigning all samples within the spatio-temporal window around that sample to this new auditory event; (3) estimating the TOA, DOA, and amplitude of the auditory event using all assigned samples; and (4) removing all corresponding samples from the SRIR. This process was repeated until no nonzero samples remained in the SRIR, and consequently, all samples in the SRIR were assigned to auditory events. The definition of the spatio-temporal window and the parameter estimation are detailed in the following section.

3.2.1 Spatio-Temporal Window

An asymmetric temporal window centered around the TOA τ was used to select contributions of the SRIR that belong to each auditory event. The TOA estimation is described in the next section. The window starts $\tau_1 = 0.5$ ms before τ to account for the pre-ringing of band-limited signals and ends $\tau_2 = 0.8$ ms after τ to model summing localization of coherent sources, i.e., the time in which the

auditory system averages incoming sound to form a single auditory event [17, chapter 3.1]. In a previous study, a value of 1 ms was used [15]. Here, it was adjusted to 0.8 ms to ensure that the very first early reflection[3] (in many cases the floor reflection) was never grouped with the direct sound, because the grouping of reflections would neglect possible comb filter effects that might occur, altering the timbre of the direct sound [9, 20], which was noticeable during informal listening.

The lateral width of the window as a function of the DOA was obtained as illustrated in Fig. 2, i.e., by interpolating empirical data from Best et al. [21] to the DOA of the auditory event. The estimation of the DOA is described in SEC. 3.2.2. The polar width was linearly interpolated from the values obtained in SEC. 2.

3.2.2 Parameterization of Auditory Events

Auditory events were each parameterized by a TOA, DOA, and amplitude. The TOA, τ, was taken as the time of the absolute maximum of p_{omni} within the spatio-temporal window. The amplitude was calculated as the RMS average of all samples within the spatio-temporal window as defined in SEC. 3.2.1:

$$a_0 = \sqrt{\frac{1}{\tau_2 + \tau_1} \int_{\tau-\tau_1}^{\tau+\tau_2} p^2(t)\,\mathrm{d}t}. \tag{2}$$

To reflect the fact that the auditory system exploits different mechanisms for localization in horizontal and median planes [17], the DOA was calculated separately for the lateral and polar angle using the weighted average,

$$\varphi_0 = \frac{1}{(\tau_2 + \tau_1) \cdot a_0^2} \int_{\tau-\tau_1}^{\tau+\tau_2} p^2(t)\varphi(t)\,\mathrm{d}t, \tag{3}$$

and the circular weighted average,

$$\theta_0 = \angle\left(\int_{\tau-\tau_1}^{\tau+\tau_2} p^2(t)\mathrm{e}^{-\mathrm{j}\theta(t)}\,\mathrm{d}t \right), \tag{4}$$

with $\angle(\,\cdot\,)$ denoting the angle of a complex number and $\mathrm{j} = \sqrt{-1}$ the imaginary unit. The weight $p^2(t)$ was chosen to approximate the level dependence of summing localization [17, chapter 3.1]. For simplicity, perfect summing localization was also assumed for the polar angle, although this assumption holds only partially [22, 38].

3.3 Masking Threshold

In the second stage of the process, a masking threshold was employed to eliminate potentially inaudible early reflections and diffuse reverberation. This threshold was implemented with both temporal and spatial dependencies, taking into account that the audibility of a reflection is influenced by its delay and spatial separation from the direct sound.

[3] This arrived shortly after 0.8 ms or later in the test cases.

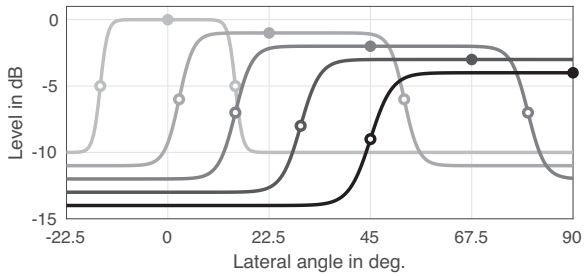

Fig. 8. Lateral dependence of the masking threshold function with a dynamic range of 10 dB for direct sound angles of $\varphi = \{0, 22.5^\circ, 45^\circ, 67.5^\circ, 90^\circ\}$ (solid dots). The circled –5 dB points, defining the window width, are taken from Fig. 2. Note that the lateral angle does not exceed 90° in the chosen coordinate system. Curves are offset in level for visibility.

The parameters of the masking threshold were tuned within ranges found in previous studies discussed in SEC. 1.2 with the goals to i) detect only a few reflections after the mixing time and ii) detect at least ten reflections up to the mixing time—including the (very) first reflections, which were deemed important for preserving the timbre of the direct sound due to comb filter effects [20, 39].

The temporal dependency used an initial threshold level of $a_0 = -17.25$ dB (average of results taken from [25, 26]) relative to the direct sound level and a decay rate of 1 dB/ms (average of results from [14, 25, 26]) starting at the direct sound TOA τ_0. Earlier studies indicated that the double-sloped curve proposed by Jensen et al. [11] depicted in Fig. 3 resulted in the improbable detection of audible reflections after the mixing time (cf. [15; 11, Fig. 10]). Therefore, the v-shaped threshold curve proposed by Olive and Toole [14] was modeled. This was done by adding 35% of the reverberant energy to the threshold. The reverberant energy was calculated as the RMS energy of the SRIR up to the current point in time, excluding the direct sound energy.

The spatial dependency of the masking threshold was realized by reducing the threshold with increasing spatial separation from the direct sound DOA (φ_0, θ_0). This was implemented as a two-dimensional window function around the DOA. The widths in the lateral and polar directions were determined according to Figs. 2 and5, respectively, and the depth was set to 10 dB (cf. SEC. 1.2), in accordance with prior studies reporting a reduced masking threshold of 6−10 dB for spatially separated sources [12, 40]. Hyperbolic tangent windows were used to achieve smooth transitions between the two levels. The final shape of the lateral dependency of the masking threshold is shown in Fig. 8 for a few example directions. The masking threshold was calculated independently in the lateral and polar direction, and the smaller value was used for further processing. This accounts for the fact that a reflection (concurrent source) becomes audible if it exceeds any of the spatial thresholds.

All reflections exceeding the masking threshold were marked audible, and all other reflections were contributed to the late reverberation described in SEC. 3.5. The full spatio-temporal evolution of the masking threshold for one test case is shown in Fig. 9. It should be noted that the

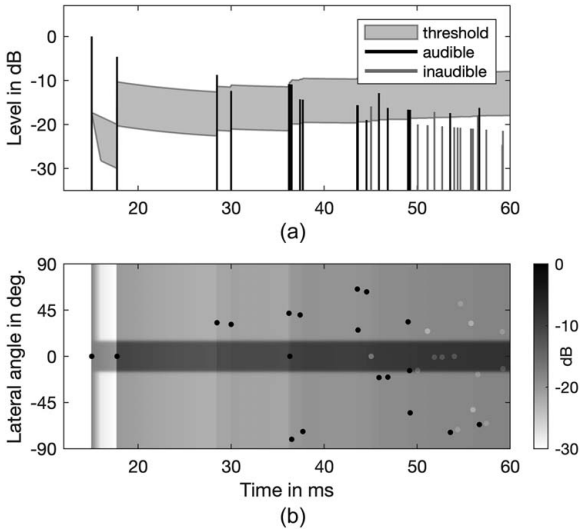

Fig. 9. (a) Temporal evolution of the masking threshold, with the gray area reflecting the possible values, depending on the direction of the incident. This visualization can be interpreted as a side view of Fig. 8 with an added time dimension. It can be seen that audible reflections raise the threshold, whereas inaudible reflections do not change its value. (b) Spatio-temporal evolution of the masking threshold with the angular dependency in the lateral direction. This visualization can be interpreted as a top view of the above plot, with the threshold level now being color-coded and the y axis showing the angular dependency instead. The black and gray dots correspond to the audible and inaudible reflections shown in the plot above. It is visible how some reflections arriving from the side are marked as audible, although they have a lower amplitude than some other inaudible reflections arriving from the median plane (lateral angle 0).

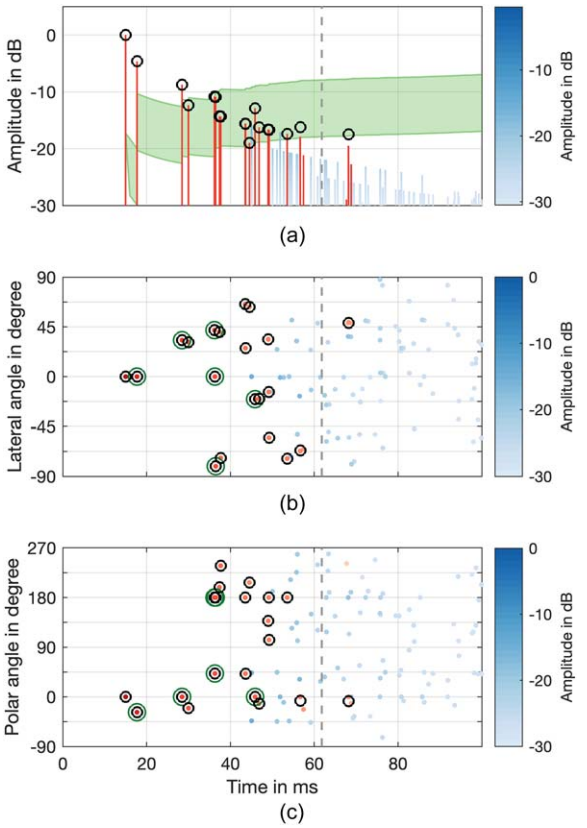

Fig. 10. SRIRs, detected early reflections and masking threshold, simulated SRIR of the medium, dry room ($T_{60} = 0.5$ s, $V = 1{,}000\,m^3$, cf. SEC. 3.1), center receiver position. RIRs without spatial information are shown in (a); (b) and (c) show the lateral and polar distribution, respectively. Potentially audible SRIR contributions are highlighted in dark red. Black circles indicate the TOA, amplitude, and DOA of detected reflections. The top row additionally shows the value range of the masking threshold functions in green. The gray dashed line indicates the perceptual mixing time. The additional green circles indicate the first-order reflections from the image source model.

conceptual distinction between early and late reflections is irrelevant to the masking threshold and was only used as a criterion for parameter tuning. Late reflections arriving after the mixing time that exceed the masking threshold are therefore treated as (early) reflections by the proposed algorithm. This can be observed in Fig. 10.

Examples of detected reflections in the empty shoebox room ($V = 1{,}000\,m^3$; RT = 0.5 s) are shown in Fig. 10. The spatio-temporal evolution of the masking threshold, calculated independently in the lateral and the polar direction, is visible in the top row of Fig. 10. The lateral dependency is best observed in the center row, where relatively loud contributions around $\varphi = 0$ were discarded for $t \gtrsim 40$ ms. The influence of the polar dependency is best observed in the bottom row, where the ceiling reflection at $\theta = 45°$ was detected. For $t \gtrsim 50$ ms, multiple SRIR samples were often grouped and detected as one reflection.

3.4 Selection of Early Reflections

In the next step, a fixed number of reflections were selected from the list extracted in the previous step to account for the available computational resources or desired degree of realism. Three simple selection methods were initially considered in a previous study: (i) Use the N *first* reflections, (ii) the N *loudest*, or (iii) the N reflections that *exceed* the masking threshold function the most [15]. The first approach had a tendency to favor early second-order reflections over louder but later-arriving first-order reflections. This also caused an imbalanced selection of reflections arriving from the left and right in the tested cases. The *exceed* method led to a more balanced selection with respect to the lateral angle but always discarded the floor reflection. The *loudest* reflections were used because this avoided these problems and is similar to the approach of Coleman et al. [1, 2].

3.5 Late Reverberation Encoding

The late reverberation was encoded from the residual RMS energy, i.e., the energy of the SRIR without the direct sound and the N selected early reflections. The residual energy was calculated for non-overlapping blocks of 256 samples at a sampling rate of 44.1 kHz. For parameter reduction, the RMS estimates in decibels were approximated by least-squares fitting of two first-order polynomials. The first polynomial starts at the direct sound TOA and ends at the TOA of the last rendered early reflection. The second

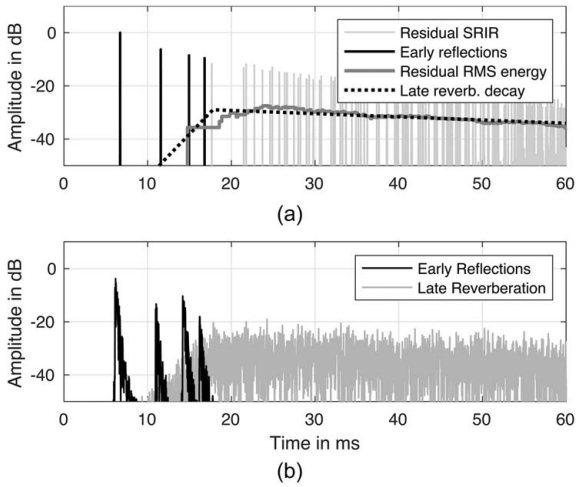

Fig. 11. Parametric rendering of the small dry room using the three loudest early reflections and double-sloped late reverberation. (a) Parametric representation of the SRIR. (b) binaural RIR (right channel).

polynomial accounts for the remainder of the SRIR. Both polynomials were designed with the constraint of equal energy at the intersection point, i.e., the position of the latest early reflection [cf. Fig. 11(a)].

3.6 Decoding

The direct sound and early reflections were rendered using HRTFs from the FABIAN HRTF database [41, 42] that were interpolated to the exact DOAs using spherical harmonics of order 35 and added to the binaural RIRs at τ_i with an amplitude of a_i. The late reverberation was modeled by Gaussian white noise with a diffuse-field interaural cross-correlation [43]. The noise was multiplied with the polynomials estimated from the RMS residual energy to achieve the desired temporal shape [cf. Fig. 11(b)]. Computationally cheaper possibilities would be to use feedback delay networks [44] or velvet noise [45]. Using multiple instances with fixed but differing reverberation times in a send-bus-like approach could further increase the performance for gaming use cases with numerous sources [4].

4 PERCEPTUAL EVALUATION

As a proof of concept, the ability of the proposed algorithm to detect and select salient early reflections was evaluated in a listening test through a comparison between the parametric renderings and the reference. A more detailed qualitative analysis of the small dry room in a previous study showed that other qualities such as the perceived tone color, source position, distance, width, and externalization correlated with the overall difference ratings [15]. To focus on the detection of early reflections and eliminate any effects of the decoding chain, the evaluation was restricted to the renderings of the shoebox rooms described in SEC. 3.1, because these reference stimuli could be generated with the same processing described in SEC. 3.6.

4.1 Listening Test Stimuli

The reference was obtained by a direct binaural rendering of the SRIRs from the ISM. This was done by applying HRTFs from the FABIAN database—the same as used for the parametric rendering—to all reflections from the ISM. The late reverberation as generated in SEC. 3.1 was binauralized as described in SEC. 3.6. This ensured that differences between the test conditions and the reference could almost exclusively be ascribed to the rendering of early reflections. Parametric renderings of the small room ($V = 200\,m^3$), including all reverberation times $T_{60} = \{0.5, 1, 2\}$ s in the center and corner receiver positions with $N = \{0, 3, 6, 9, \text{all}\}$ audible reflections (i.e., the N loudest reflections that were marked audible by the masking threshold) were chosen as test conditions. The RMS levels of the test conditions were adjusted to the reference to exclude loudness as a cue. The loudness across the rooms was adjusted using RMS normalization. Anechoic male speech was used as audio content (first 5 s from track 50 of the EBU SQUAM CD[4]).

4.2 Study Protocol

Forty subjects participated in the listening test (eight female, 31 male, one unspecified, mean age 30 years) with an average of 3.1 h of audio-related tasks per day. Out of these participants, 26 had previously taken part in listening tests.

The experiment was conducted online over a web interface [46] with a modified version of the MUSHRA method [47]. The participants were first asked to set the volume of the headphones to match the loudness of the test stimuli to that of a male person speaking at a distance of 3 m. After an introduction to the user interface, a brief training was given to familiarize the subjects with the rating procedure. The training contained an exemplary stimulus and the corresponding references to cover the range of differences to be expected during the test. The subjects were then asked to rate the differences and were instructed to rely on their own interpretation of "large" differences but to remain consistent throughout the test. The subjects were told to take their time at will and to listen to the stimuli as often as, and in any order they wanted.

The presentation order of reverberation times and receiver positions was randomized, resulting in six ratings per screen (*N loudest*, *all audible*, and *hidden reference*; also in randomized order). The stimuli for each of the six used room/receiver configurations were presented on a separate page. The subjects spent an average of 17 min on the test (not including the training procedure).

4.3 Analysis and Results

Fig. 12 shows the median ratings and 95% bootstrapped confidence intervals (non-parametric resampling, bias-corrected, and accelerated calculation). The median ratings across all rooms are $\tilde{\mu} = \{0.4, 0.34, 0.25, 0.22, 0.15, 0\}$ for $N = \{0, 3, 6, 12, \text{all}, \text{ref}\}$. Because of the ratings not being normally distributed, multilevel models, which only

[4] https://tech.ebu.ch/publications/sqamcd.

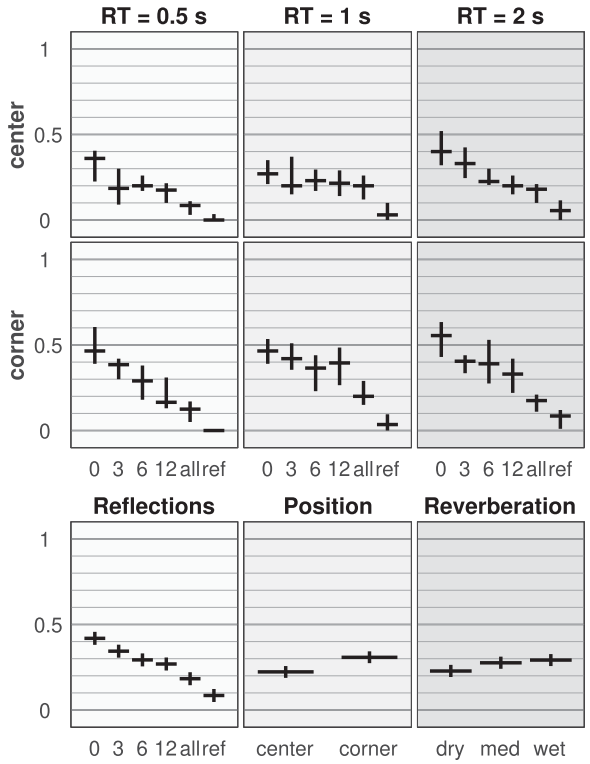

Fig. 12. Results of the perceptual evaluation. (a) *Difference* ratings from the listening test given by the median and 95% bootstrapped confidence intervals. Rows and columns indicate the receiver position and reverberation time. *Ref* denotes the hidden reference, *all* denotes all early reflections that are marked *audible* by the masking threshold (cf. SEC. 3.3). {0, 3, 6, 12} denotes renderings containing only the N loudest *audible* reflections. (b–d) Estimated marginal means and 95% confidence intervals of the *Difference* ratings for the factor levels *Number of Reflections* (b), *Receiver Position* (c), and *Reverberation* (d).

require normally distributed residuals [48], were used for the statistical analysis. The model for *difference* accounts for $R^2 = 50\%$ of the variance (marginal $R^2 = 31\%$ [49]) and the main effects of the three factors *(number of) reflections*, *receiver position*, and *reverberation* were determined to be statistically significant ($p < 0.001$).

The *reflections* have the largest effect, and differences clearly decrease with increasing N (estimated marginal means $\hat{\mu} = \{0.42, 0.34, 0.29, 0.27, 0.18, 0.09\}$ for $N = \{0, 3, 6, 12, \text{all}, \text{ref}\}$), which accounts for 34% of the variance [50, Eq. (20.30)]. Difference ratings are lower in the center than in the corner receiver position ($\hat{\mu} = \{0.22, 0.31\}$ for {center, corner}), accounting for 9.7% of the variance. In the dry room, the difference was rated lower than in the medium and wet room ($\hat{\mu} = \{0.23, 0.28, 0.29\}$ for RT $= \{0.5, 1, 2\}$ s), accounting for 4.8% of the variance.

Dunn-Šidák corrected pairwise comparisons showed statistically significant differences ($p < 0.001$) between almost all levels of each factor. The only exceptions are the ratings of $N = 6$ compared to $N = 12$ [$p = 0.812$, cf. Fig. 12(b)] and the ratings of the medium and wet room [$p = 0.369$, cf. Fig. 12(d)].

Additionally, first-order interactions show that for $N = \{0, 3, 6, 12\}$, the ratings are higher in the corner than in the center receiver position. In contrast, the listener position does not influence the ratings for *all audible* and the hidden reference.

Notably, some subjects failed to correctly identify the reference and gave non-zero ratings for some conditions. The statistical analysis was run with and without excluding those subjects to assess their effect on the results. Because the general findings (significant effects and effect sizes) were almost identical in both cases, all subjects were included in the above analysis to improve its robustness.

5 DISCUSSION

This section discusses the performance of the reflection detection and selection algorithm on simulated SRIRs and the results of the perceptual evaluation.

5.1 Physical Evaluation

Applying the masking threshold reduced the number of reflections drastically. In the small rooms, the masking threshold detected 24–42 *audible* reflections. This is a reduction of 63%–81% considering the 115–129 image sources before the perceptual mixing time [35]. In the medium and large rooms, the reduction rates were lower (5%–53% in the medium rooms, 32%–84% in the large rooms). The lower reduction rates were mainly due to a significantly smaller number of image sources before the mixing time (38–45 in the medium and large rooms) and, thus, fewer inaudible reflections. Also, in those rooms, multiple reflections were often being summed to auditory events whose energies exceeded the masking threshold [an example of this can be seen after the mixing time in Fig. 10(a)]. The summing of multiple reflections into auditory events is not entirely preventable and happens more often the closer the reflections are in time, so the masking threshold should be tuned further to suppress these events after the mixing time.

The polar dependency of the masking threshold helped select the ceiling and floor reflections more often. The floor reflection was determined to be audible in all rooms, and the ceiling reflection was detected in most cases. It was discarded in the medium room in the corner position and in the large room with medium reverberation time. In these cases, the opening angle between the direct sound and ceiling reflection was below the polar threshold for source separation ($\theta = 37.5^\circ$ in the medium room and $\theta = 38^\circ$ in the large room). Reflections from the front wall (as viewed from the receiver, cf. Fig. 7) were suppressed in all cases for the wall receiver position, where the front wall reflection arrives from the same direction as the direct sound. In the other cases, both front and back wall reflections were detected.

When the receiver was positioned close to the corners of the room facing away from the walls, the polar dependency of the masking threshold helped to detect strong and early second-order reflections arriving from behind the listener. In informal listening, this improved the perceived auditory impression compared to cases with the threshold's polar dependency disabled. Moreover, rendering of the six *loud-*

est audible reflections in these cases yielded better auditory impressions than rendering the six first-order reflections.

5.2 Perceptual Evaluation

The focus of the perceptual evaluation was to test if the detection and selection algorithm could accurately simulate the sound field within a room for VR/AR applications in the selected rooms. Two main questions were asked: (1) Does removing reflections based on the masking threshold cause audible differences? (2) If a fixed number of reflections are rendered, how does the number of reflections affect the perceived differences?

The results of the study showed that applying the masking threshold does cause small but audible differences [c.f. Fig. 12(b), condition *all*]. Out of the 40 participants, 14 reported coloration as one of the primary cues. However, the coloration was generally low and most likely only detectable in direct comparison to the reference. The problem may thus be of little relevance for many VR and AR applications. This is supported by small differences between the estimated marginal means for *all audible* reflections (0.18) and the *hidden reference* (0.09). Because each reflection adds a comb filter to the room transfer function, it is plausible that removing and grouping reflections can create comb filter-like coloration. The proposed algorithm aims to remove only inaudible reflections to avoid this problem, but it may still cause audible differences when removing multiple reflections that each would be inaudible.

The influence of single parameters should be investigated in more detail to account for this possibility. To preserve coloration, the time dependency of the masking threshold was tuned to include the first few early reflections because they can cause larger distinct comb filter effects than later reflections. Comb-filter audibility thresholds vary depending on the stimuli, with the lowest thresholds for transient-rich signals occurring at around 0.5–2 ms delay and for speech stimuli at delays of up to 15 ms [39, 20].

The findings presented in Fig. 12(b) show that a relatively small number of reflections could provide a good perceptual approximation of the sound field within the tested rooms. The perceived differences decrease monotonically with an increasing number of rendered reflections. The results show that within the scope of the study, the perceptual effect of rendering six reflections compared to none is approximately equally large as rendering all detected reflections instead of six. The authors thus consider the final choice of how many reflections shall be rendered to be an application-depending trade-off between available computational resources and the desired audio quality. The perceptual cost of omitting reflections also depends on the audio content, with less strong reflections being more noticeable for transient-rich sounds like castanets and more subtle tonal changes being noticeable when using white noise.

The smallest effect on perceived differences was observed for the reverberation time, suggesting that the proposed encoding works almost equally well regardless of the reverberation. However, a previous study found that perceived differences slightly decreased with increasing reverberation [15], so more reliable results for this condition may be obtained by directly comparing stimuli for varying degrees of reverberation but the same number of rendered reflections on a single rating screen.

It is worth noting that ratings for the hidden reference condition displayed small deviations from zero. Some participants assigned large differences even when they reported that the differences were barely audible during the open-ended questions at the end of the test. It is possible that participants believed that failing to report differences would reflect poorly on their performance, leading to an exaggerated rating of differences, or that factors such as the background noise and playback level that could not be fully controlled during the online experiment caused these issues.

To avoid this, an anchor could have been used, and the subjects could have been instructed to always use the entire scale. However, in this case, it was challenging to identify an appropriate anchor because of the complex and nuanced nature of the stimuli being tested. To this end, the authors considered the condition that contained only the direct sound followed by the diffuse late reverberation to be the most suitable and meaningful anchor-like stimulus. The authors intentionally avoided using the entire rating scale to preserve potential differences between test conditions across multiple rating screens. As such, a multilevel model was chosen to be employed for the statistical analysis, which can partially account for the different rating behavior of the subjects by means of the estimated random intercepts.

6 FUTURE WORK

So far, the perceptual evaluation was restricted to the overall perceived difference between renderings, focusing on the detection and selection of reflections rather than other parts of the encoding and decoding chain. The influence of the windowing and masking threshold parameters on the auditory impression remains to be investigated in a detailed analysis. For instance, it will be of interest to see if the results for the polar threshold for median plane source separation derived in SEC. 2 can be validated in a listening test.

At this development stage, the encoding did not consider frequency-dependent rendering, directional sound sources, directional late reverberation, and diffuse reflections. Whereas the current encoding of early reflections would already be able to account for directional sources, the other aspects would require special processing. Frequency-dependent rendering could be achieved by estimating reflection filters as proposed by Arend et al. [51] in combination with a frequency-dependent analysis and reproduction of the late reverberation (cf. [1, 4, 52, 53]).

Diffuse reflections could be considered by adjusting the masking threshold, taking into account increased sensitivity of up to 8 dB for diffuse reflections [40]. The diffuseness of a reflection might be assessed by comparing its temporal structure against an ideal band-limited pulse in combination with an analysis of the variance in the sample-wise DOA estimate of the reflection. Only after these factors are con-

sidered would it be sensible to test the suggested approach on more complex input data like band-limited simulations or acoustically measured data. Although the general applicability of the encoding stage to band-limited input was shown in Brinkmann et al. 15, Fig. 3, left], it might be expected that the image-source–based simulations used in this study represent the best possible input data and thus might be considered an upper performance limit.

The effect of strong (late) reflections that are perceived as echoes could be accounted for by using an additional echo threshold during the encoding stage. Such a threshold is already implemented but was not discussed here for brevity [15]. For reflections exceeding the echo threshold, the calculation of the masking threshold would reset from the DOA and the amplitude of the direct sound to that of the echo from that point on.

Listener translation could be implemented similarly to Arend [51] or Müller and Zotter [8, 54]. However, translation and head rotations are challenging because they change the DOA of the direct sound and reflections, thus constantly updating the masking threshold and, consequently, the selected reflections. This might be solved by moving the spatial dependency of the masking threshold from the encoding to the decoding stage, which, as a side effect, would cause more detected reflections to be stored in the parametric SRIR.

For scenarios in which the source and listener may translate, parameters must be encoded offline for numerous spatial source-listener location pairs in large scenes [4]. This poses additional challenges regarding spatial smoothness. A previous investigation in a nonempty room suggests that the extracted parameters vary smoothly over space for the most part, but discontinuities necessarily occur when a reflection's level crosses the threshold function [15]. Although the perceptual evaluation indicates audible effects if excluding all *inaudible* reflections, the reflection would pass the threshold individually, which potentially mitigates artifacts related to activating and deactivating single reflections. Considering absolute sound levels as Green and Kahle [55] did will be difficult, assuming that the playback level will be user-controlled in most cases.

7 CONCLUSION

The authors propose a parametric encoding of SRIRs with a focus on detecting and selecting perceptually salient early reflections. In principle, it can be applied to any SRIR that includes DOA information. The encoding technique utilizes a novel spatio-temporal windowing method for segmenting SRIRs into auditory events, which are then parameterized. Salient early reflections are selected using perceptually motivated masking thresholds. The proposed encoding was evaluated against reference simulations obtained using image sources and stochastic late reverberation. Applying the reflection selection algorithm produced minor differences, which were noticeable in direct comparison. The perceptual transparency of the encoding and decoding chain was not evaluated as a whole.

8 ACKNOWLEDGMENT

The authors would like to thank the participants of the listening test for their time.

9 REFERENCES

[1] P. Coleman, A. Franck, P. J. B. Jackson, et al., "Object-Based Reverberation for Spatial Audio," *J. Audio Eng. Soc.*, vol. 65, no. 1/2, pp. 66–77 (2017 Feb.). https://doi.org/10.17743/jaes.2016.0059.

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

[3] P. Stade, J. 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* (2017 May), paper 9688.

[4] N. Raghuvanshi and J. Snyder, "Parametric Directional Coding for Precomputed Sound Propagation," *ACM Trans. Graph.*, vol. 37, no. 4, paper 108 (2018 Aug.). https://doi.org/10.1145/3197517.3201339.

[5] 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 Jul.). https://doi.org/10.17743/jaes.2021.0009.

[6] L. Remaggi, P. J. B. Jackson, P. Coleman, and W. Wang, "Acoustic Reflector Localization: Novel Image Source Reversion and Direct Localization Methods," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 2, pp. 296–309 (2017 Feb.). https://doi.org/10.1109/TASLP.2016.2633802.

[7] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of Glottal Closure Instants in Voiced Speech Using the DYPSA Algorithm," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 1, pp. 34–43 (2007 Jan.). https://doi.org/10.1109/TASL.2006.876878.

[8] K. Müller and F. Zotter, "The PerspectiveLiberator – An Upmixing 6DoF Rendering Plugin for Single-Perspective Ambisonic Room Impulse Responses," in *Proceedings of the Fortschritte Der Akustik (DAGA)*, pp. 306–309 (Vienna, Austria) (2021 Aug.).

[9] S. Bech, "Timbral Aspects of Reproduced Sound in Small Rooms. I," *J. Acoust. Soc. Am.*, vol. 97, no. 3, pp. 1717–1726 (1995 Mar.). https://doi.org/10.1121/1.413047.

[10] S. Bech, "Spatial Aspects of Reproduced Sound in Small Rooms," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 434–445 (1998 Jan.). https://doi.org/10.1121/1.421098.

[11] R. E. Jensen and T. S. Welti, "The Importance of Reflections in a Binaural Room Impulse Response," presented at the *114th Convention of the Audio Engineering Society* (2003 Mar.), paper 5839.

[12] J. M. Buchholz, J. Mourjopoulos, and J. Blauert, "Room Masking: Understanding and Modelling the Masking of Reflections in Rooms," presented at the *110th Con-*

vention of Audio Engineering Society (2001 May), paper 5403.

[13] M. Röhrbein and A. Lindau, "Reducing the Temporal Resolution of Spatial Impulse Responses With an Auditory Model," in *Proceedings of the Fortschritte Der Akustik (DAGA)*, pp. 327–328 (Düsseldorf, Germany) (2011 Mar.).

[14] S. E. Olive and F. E. Toole, "The Detection of Reflections in Typical Rooms," *J. Audio Eng. Soc.*, vol. 37, no. 7/8, pp. 539–553 (1989 Jul.).

[15] 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* (2020 May), paper 10380.

[16] Deutsches Institut fur Normung, "Acoustics - Terminology," Tech. Rep. DIN 1320:2009-12 (2009 Dec.). https://doi.org/10.31030/1544140.

[17] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization (Revised Edition)* (MIT Press, Cambridge, MA, 1997). https://doi.org/10.7551/mitpress/6391.001.0001.

[18] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The Precedence Effect," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654 (1999 Oct.). https://doi.org/10.1121/1.427914.

[19] A. D. Brown, G. C. Stecker, and D. J. Tollin, "The Precedence Effect in Sound Localization," *J. Assoc. Res. Otolaryngol.*, vol. 16, no. 1, pp. 1–28 (2015 Feb.). https://doi.org/10.1007/s10162-014-0496-2.

[20] S. Brunner, H.-J. Maempel, and S. Weinzierl, "On the Audibility of Comb-Filter Distortions," presented at the *122nd Convention of the Audio Engineering Society* (2007 month), paper 7047.

[21] V. Best, A. van Schaik, and S. Carlile, "Separation of Concurrent Broadband Sound Sources by Human Listeners," *J. Acoust. Soc. Am.*, vol. 115, no. 1, pp. 324–336 (2004 Jan.). https://doi.org/10.1121/1.1632484.

[22] P. Bremen, M. M. van Wanrooij, and A. J. van Opstal, "Pinna Cues Determine Orienting Response Modes to Synchronous Sounds in Elevation," *J. Neurosci.*, vol. 30, no. 1, pp. 194–204 (2010 Jan.). https://doi.org/10.1523/JNEUROSCI.2982-09.2010.

[23] V. Pulkki, H. Pöntynen, and O. Santala, "Spatial Perception of Sound Source Distribution in the Median Plane," *J. Audio Eng. Soc.*, vol. 67, no. 11, pp. 855–870 (2019 Nov.). https://doi.org/10.17743/jaes.2019.0033.

[24] B. C. J. Moore, "Characterization of Simultaneous, Forward and Backward Masking," presented at the *12th Conference of the Audio Engineering Society* (1993 Jun.), paper 12.

[25] B. Rakerd, W. M. Hartmann, and J. Hsu, "Echo Suppression in the Horizontal and Median Sagittal Planes," *J. Acoust. Soc. Am.*, vol. 107, no. 2, pp. 1061–1064 (2000 Feb.). https://doi.org/10.1121/1.428287.

[26] D. R. Begault, B. U. McClain, and M. R. Anderson, "Early Reflection Thresholds for Anechoic and Reverberant Stimuli Within a 3-D Sound Display," in *Proceedings of the 18th International Congress on Acoustics (ICA)* (Kyoto, Japan) pp. 1267–1270 (2004 Apr.).

[27] F. Melchior, F. Gries, U. Heusinger, and J. Liebetrau, "On Early Reflection Thresholds in the Median Plane," in *Proceedings of the Fortschritte Der Akustik (DAGA)*, pp. 225–226 (Berlin, Germany) (2010 Jan.). https://doi.org/10.13140/2.1.3668.6729.

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

[29] P. Majdak, C. Hollomey, and R. Baumgartner, "AMT 1.x: A Toolbox for Reproducible Research in Auditory Modeling," *Acta Acust.*, vol. 6, paper 19 (2022 May). https://doi.org/10.1051/aacus/2022011.

[30] F. Brinkmann, M. Dinakaran, R. Pelzer, et al., "The HUTUBS HRTF Database," *FG Audiokommunikation* (2019 May). https://doi.org/10.14279/depositonce-8487.

[31] O. Pele and M. Werman, "Fast and Robust Earth Mover's Distances," in *Proceedings of the IEEE 12th International Conference on Computer Vision*, pp. 460–467 (Kyoto, Japan) (2009 Sep.). https://doi.org/10.1109/ICCV.2009.5459199.

[32] A. Nieslony, "Random Numbers From a User Defined Distribution," MATLAB Central File Exchange, https://de.mathworks.com/matlabcentral/fileexchange/26003-random-numbers-from-a-user-defined-distribution (accessed Aug. 1, 2020).

[33] N. Kolbe, "Compact Matlab Code for the Computation of the 1- and 2-Wasserstein Distances in 1D," https://github.com/nklb/wasserstein-distance (2020 Oct.).

[34] F. Brinkmann and S. Weinzierl, "AKtools – An Open Software Toolbox for Signal Acquisition, Processing, and Inspection in Acoustics," presented at the *142nd Convention of the Audio Engineering Society* (2017 May), e-Brief 309.

[35] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual Evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 60, no. 11, pp. 887–898 (2012 Nov.).

[36] H. Kuttruff, *Room Acoustics* (CRC Press, London, UK, 2009), 5th ed. https://doi.org/10.1201/9781482266450.

[37] H. P. Tukuljac, V. Pulkki, H. Gamper, K. Godin, I. J. Tashev, and N. Raghuvanshi, "A Sparsity Measure for Echo Density Growth in General Environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 226–230 (Brighton, UK) (2019 May). https://doi.org/10.1109/ICASSP.2019.8682878.

[38] R. Ege, A. J. van Opstal, P. Bremen, and M. M. van Wanrooij, "Testing the Precedence Effect in the Median Plane Reveals Backward Spatial Masking of Sound," *Sci. Rep.*, vol. 8, paper 8670 (2018 Jun.). https://doi.org/10.1038/s41598-018-26834-2.

[39] T. Anazawa, Y. Takahashi, and A. H. Clegg, "Digital Time-Coherent Recording Technique," presented at the *83rd Convention of the Audio Engineering Society* (1987 Oct.), paper 2493.

[40] F. Wendt and R. Höldrich, "Precedence Effect for Specular and Diffuse Reflections," *Acta Acust.*, vol. 5, paper 1 (2020 Dec.). https://doi.org/10.1051/aacus/2020027.

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

[42] F. Brinkmann, A. Lindau, S. Weinzierl, et al., "The FABIAN Head-Related Transfer Function Data Base," *FG Audiokommunikation* (2017 Feb.). https://doi.org/10.14279/depositonce-5718.5.

[43] C. Borß and R. Martin, "An Improved Parametric Model for Perception-Based Design of Virtual Acoustics," in *Proceedings of the AES 35th International Conference* (2009 Feb.), paper 3.

[44] H. Steffens, S. van de Par, and S. D. Ewert, "Perceptual Relevance of Speaker Directivity Modelling in Virtual Rooms," in *Proceedings of the 23rd International Congress on Acoustics*, pp. 2651–2658 (Aachen, Germany) (2019 Sep.). https://doi.org/10.18154/RWTH-CONV-239630.

[45] V. Välimäki, B. Holm-Rasmussen, B. Alary, and H.-M. Lehtonen, "Late Reverberation Synthesis Using Filtered Velvet Noise," *Appl. Sci.*, vol. 7, no. 5, paper 483 (2017 May). https://doi.org/10.3390/app7050483.

[46] M. Schoeffler, S. Bartoscheck, F.-R. Stöter, et al., "webMUSHRA — A Comprehensive Framework for Web-Based Listening Tests," *J. Open Res. Softw.*, vol. 6, no. 1, paper 8 (2018 Feb.). https://doi.org/10.5334/jors.187.

[47] International Telecommunication Union, "Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems," *Recommendation ITU-R BS.1534-3* (2015 Oct.). https://www.itu.int/rec/R-REC-BS.1534-3-201510-I/.

[48] J. J. Hox, *Multilevel Analysis: Techniques and Applications*, Quantitative Methodology Series (Routledge, New York, NY, 2010), 2nd ed. https://doi.org/10.4324/9780203852279.

[49] S. Nakagawa and H. Schielzeth, "A General and Simple Method for Obtaining R^2 From Generalized Linear Mixed-Effects Models," *Methods Ecol. Evol.*, vol. 4, no. 2, pp. 133–142 (2012 Dec.). https://doi.org/10.1111/j.2041-210x.2012.00261.x.

[50] M. Eid, M. Gollwitzer, and M. Schmitt, *Statistik Und Forschungsmethoden* (Beltz, Weinheim, Germany, 2017), 5th ed.

[51] 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 Jul.). https://doi.org/10.17743/jaes.2021.0009.

[52] P. Stade, J. Arend, and C. Pörschmann, "Perceptual Evaluation of Synthetic Late Binaural Reverberation Based on a Parametric Model," in *Proceedings of the AES International Conference on Headphone Technology* (2016 Aug.), paper 3-3.

[53] B. Alary, A. Politis, S. Schlecht, and V. Välimäki, "Directional Feedback Delay Network," *J. Audio Eng. Soc.*, vol. 67, no. 10, pp. 752–762 (2019 Oct.). https://doi.org/10.17743/jaes.2019.0026.

[54] K. Müller and F. Zotter, "Auralization Based on Multi-Perspective Ambisonic Room Impulse Responses," *Acta Acust.*, vol. 4, no. 6, paper 25 (2020 Nov.). https://doi.org/10.1051/aacus/2020024.

[55] E. Green and E. Kahle, "Dynamic Spatial Responsiveness in Concert Halls," *Acoustics*, vol. 1, no. 3, pp. 549–560 (2019 Jul.). https://doi.org/10.3390/acoustics1030031.

THE AUTHORS


Tobias Jüterbock


Fabian Brinkmann


Hannes Gamper


Nikunj Raghuvanshi


Stefan Weinzierl

Tobias Jüterbock is a research associate at the Technical University of Berlin, Germany, where he received his M.Sc. degree in physical engineering in 2022. His interests are virtual acoustics, room acoustics, electroacoustics, and signal processing. Currently, he is involved in developing differentiable ray tracing techniques for room acoustics.

•

Fabian Brinkmann received an M.A. degree in Communication Sciences and Technical Acoustics in 2011 and Dr. rer. nat. degree in 2019 from the Technical University of Berlin, Germany. He focuses on the fields of signal processing and evaluation approaches for spatial audio. Fabian is a member of the AES, German Acoustical Society (DEGA), and the European Acoustics Association (EAA) technical committee for psychological and physiological acoustics.

•

Hannes Gamper is a researcher in the Audio and Acoustics Research Group at Microsoft Research in Redmond. He received his Ph.D. degree from Aalto University in Finland. At Microsoft, he mainly works on projects related to spatial sound and psychoacoustics and has contributed to products including HoloLens and Windows 10.

•

Nikunj Raghuvanshi leads research projects at Microsoft Research's Redmond labs in the areas of spatial audio, computational acoustics, and computer graphics for games and AR/VR. As a senior researcher, he's tasked with conceiving new research directions, working on technical problems with collaborators, mentoring interns, publishing findings, giving talks, and engaging closely with engineering groups to translate the ideas into real-world impact. Over the last decade, he has led project Triton, a first-of-its-kind wave acoustics system that is now in production use in multiple major Microsoft products such as Gears of War and Windows 10. Triton is currently in the process of being opened up for external use as part of Project Acoustics. Nikunj has published and given talks at top academic and industrial venues across disciplines: ACM SIGGRAPH, Audio Engineering Society, Acoustical Society of America, and Game Developers Conference, and he served on the SIGGRAPH papers committee. Before Microsoft, he did his Ph.D. studies at UNC Chapel Hill, where his work helped initiate sound as a new research direction. His entire thesis code was licensed by Microsoft.

•

Stefan Weinzierl is head of the Audio Communication Group at the Technische Universität Berlin. His research is focused on audio technology, virtual acoustics, room acoustics, and musical acoustics. With a diploma in physics and sound engineering, he received his Ph.D. in musical acoustics. He is coordinating a Master's program in Audio Communication and Technology at TU Berlin and has coordinated international research consortia in the field of virtual acoustics (SEACEN) and music information retrieval (ABC_DJ).