Audio Engineering Society

Convention Paper

Presented at the 148th Convention
2020 May 25 – 28, Vienna, Austria

Towards encoding perceptually salient early reflections for parametric spatial audio rendering

Fabian Brinkmann[1*], Hannes Gamper[2], Nikunj Raghuvanshi[2], and Ivan Tashev[2]

[1]*Audio Communication Group, Technical University of Berlin, Germany.*
[2]*Microsoft Research Redmond, WA, USA.*
[*]*The work was done as a research intern at Microsoft Research Labs in Redmond, WA, USA.*

Correspondence should be addressed to Fabian Brinkmann (fabian.brinkmann@tu-berlin.de)

ABSTRACT

Parametric spatial audio rendering promises fast and perceptually convincing audio cues that remain playback-system agnostic and enable aesthetic modifications of the acoustic experience within games and virtual reality. We propose a parametric encoder for spatial room impulse responses that is tested with nine simulated rooms spanning a large range of sizes and reverberation times. A key component of the pipeline is a perceptually inspired model for determining a minimal set of salient early reflections to reduce computational complexity. The results of a listening study with 27 subjects suggest that rendering six early reflections is indiscernible from a fully-rendered reference for the tested speech content and frequency-independent room simulations based on the image source method. However, the proposed model requires further improvements with respect to detecting and selecting the most-salient early reflections.

1 Introduction

Virtual reality (VR), mixed reality and gaming applications must perform 3-D sound rendering within a small fraction of a single CPU core since 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 and provide consistent audio-visual cues to 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 3-D sound in real time. Common parametric models include various aspects of 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, 2, 3, 4, 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.

Past studies used either listening tests [6, 7, 8] or binaural models [9, 10] to determine and in some cases discard inaudible reflections in rooms. 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. However, previous work found that for speech and music content 5–11 reflections may be sufficient for perceptually

transparent rendering [10], which is promising.

On the other hand, prior methods that directly encode SRIRs use a fixed number of reflections without explicitly determining the minimal number required for perceptual transparency. Coleman et al. encoded the six loudest reflections detected from SRIRs captured with 48 microphones [1] and 20 reflections if using four microphones [2]. Stade et al. [3] used between 50 and 200 reflections encoded from 1202 microphones. 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.

Here we propose an encoding algorithm to detect perceptually salient early reflections and evaluate it with respect to the required number of reflections. A possible use case is the automatic offline encoding of billions of spatially distributed listener and source position pairs [4]. Our results suggest that the six first-order reflections are sufficient for empty shoebox rooms. However, the algorithm needs refinement to select the perceptually most important early reflections and for the results to be more consistent across rooms.

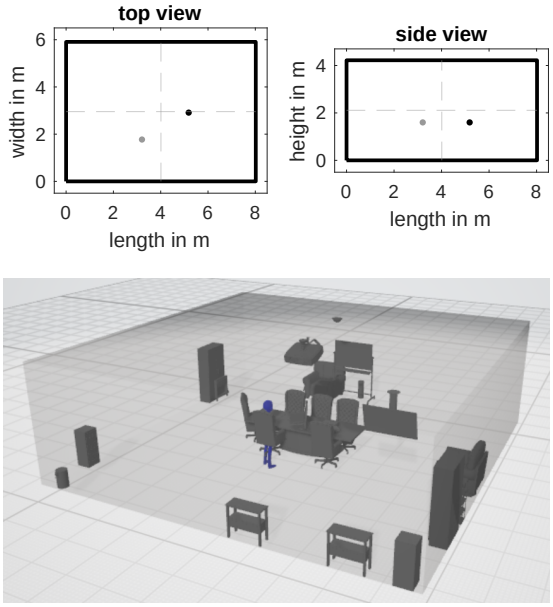

Fig. 1: Top: Sketches of the small *ISM* room including source (gray), receiver (black), and symmetry axes (dashed). The setups in the medium and large room were similar. Bottom: Picture of the *TRI* meeting room.

2 Methods

This section describes the SRIR generation, encoding and decoding. Sample code, the used SRIR data, and auralizations are available online[1].

2.1 Spatial room impulse responses

Two different SRIR data sets were generated. The *ISM* data set was computed based on a hybrid room acoustical simulation using image sources for the early reflections and stochastic decaying noise for the late reverberation [11]. This data set served as a reference for the perceptual evaluation to assure that differences between the *ISM* set and the parametric approach can almost exclusively be attributed to differences in the rendering of early reflections. The *TRI* data set was generated with the wave-based acoustic simulation *Triton* [4] to include a test case with unknown directional information. For the sake of simplicity, all simulations used frequency-independent boundary reflectivity and omnidirectional sources.

[1] https://github.com/microsoft/Perceptual_saliency_of_early_reflections

The hybrid *ISM* model was used to generate SRIRs for nine shoebox-shaped empty rooms for all combinations of room volumes $\mathrm{V} = \{200, 1000, 5000\}$ m^3 and reverberation times $\mathrm{RT} = \{0.5, 1, 2\}$ 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 RT. 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 [12]. The late reverberation was modeled as decaying white Gaussian noise and the sample-wise DOA was drawn from a uniform random distribution. To achieve a smooth transition between the early and late part, an exponential fade-in was applied to the late reverberation that started at the position of the direct sound with a level of -60 dB with respect to the level at $1.5\,\mathrm{T_{mix}}$. Sources and receivers were positioned at a height of 1.6 m, away from the rooms' symmetry axes. Their distance was two times the critical distance $d_c \approx 0.057\sqrt{\mathrm{V}/\mathrm{RT}}$ [13, Eq. 5.39] with respect to $\mathrm{RT} = 0.5$ s, i.e., the distance remained constant across RT but changed with V (cf.

Fig. 1).

For comparison, Triton was used to simulate the small room from the first database (V = 200 m^3, RT = $\{0.5, 1, 2\}$ s), as well as a small meeting room to include a case with scattering and diffraction (V = 225 m^3, RT = 0.9 s, cf. Fig. 1). The usable upper frequency limit of the simulation was 8 kHz, with a simulation sample rate of 48 kHz. The sample-wise DOA was calculated based on the intensity $\mathbf{I} = p\mathbf{v}$, where the sound velocity $\mathbf{v}$ was calculated from discrete derivatives of the simulated sound pressure p [4]. A 10th-order zero-phase Butterworth low-pass filter with a cut-off frequency of 2 kHz was applied to $\mathbf{v}$ to limit spatial fluctuation, or ringing, in the resulting DOA.

2.2 Direct-sound encoding

The onset of the direct sound was estimated using a first-moment onset detector based on the cumulative energy of the pressure response p, which proved to provide spatially smooth estimates for large numbers of source/receiver positions [4, Eq. 15]. The direct-sound TOA, τ_0, was then taken as the absolute maximum of p in a search range of 1 ms starting at the direct-sound onset. An asymmetric temporal window centered around τ_0 was used to select contributions of p that belong to the direct sound. The window starts 0.5 ms before τ_0 to account for pre-ringing of band-limited signals and ends 1 ms after τ_0 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 [14, Chap. 3.1]. From this window, the amplitude was calculated as the root-mean-squared average

$$a_0 = \sqrt{\frac{1}{1.5} \int_{\tau_0-0.5}^{\tau_0+1} p^2(t)\mathrm{d}t}\,. \tag{1}$$

To reflect the fact that the auditory system exploits different mechanisms for localization in horizontal and median planes [14], the DOA was calculated separately for the lateral angle $-90 \leq \phi \leq 90$ and polar angle $-90 \leq \theta \leq 270$ of an interaural-polar coordinate system (cf. [15, Fig. 2, right]) using the weighted average

$$\phi_0 = \frac{1}{1.5\,a_0^2} \int_{\tau_0-0.5}^{\tau_0+1} p^2(t)\phi(t)\mathrm{d}t \tag{2}$$

and the circular weighted average

$$\theta_0 = \angle \left(\int_{\tau_0-0.5}^{\tau_0+1} p^2(t)\mathrm{e}^{-j\theta(t)}\mathrm{d}t \right) \tag{3}$$

with $\angle(\cdot)$ denoting the angle of a complex number and $j = \sqrt{-1}$ the imaginary unit. The weight $p^2(t)$ was chosen to approximate the level dependence of summing localization [14, Chap. 3.1]. For simplicity, perfect summing localization was also assumed for the polar angle, although this assumption holds only partially [16].

2.3 Early reflections encoding

Perceptually salient early reflections were encoded in a three step procedure described in the following.

2.3.1 Segmentation of SRIR into reflections

First, the early part of the SRIR ($t \leq 200$ ms) was segmented into reflections. This can be interpreted as a transformation from a physical sample-based representation of *sound events* to a perceptual representation where a reflection is an *auditory event* described by its TOA, DOA, and level and may contain one or more SRIR samples. The segmentation was done iteratively by finding the sample with the largest absolute value that is not yet assigned to a reflection. The TOA τ_i of the ith reflection was given by the position of the sample, while the amplitude a_i and DOA $\sphericalangle_i$ were calculated around τ_i similar to Equations (1)–(3). However, an additional spatial window dependent on the lateral angle of the current maximum was applied to account for the ability of the auditory system to perceive multiple sources simultaneously (cf. Fig. 2, top). The width of the window was estimated visually from Best et al. [17, Fig. 3(e)] and linearly interpolated to obtain intermediate values (cf. Fig. 2, top). Values outside the spatial window were not considered for calculating a_i and $\sphericalangle_i$.

2.3.2 Detection of early reflections

In the second step, a masking threshold as a function of time, direct-sound lateral angle, and reverberant energy was used to pre-select potentially audible contributions (cf. Fig. 2, bottom). The parameters of the masking threshold were iteratively adjusted within common ranges [18] and with the goal to detect at least 10

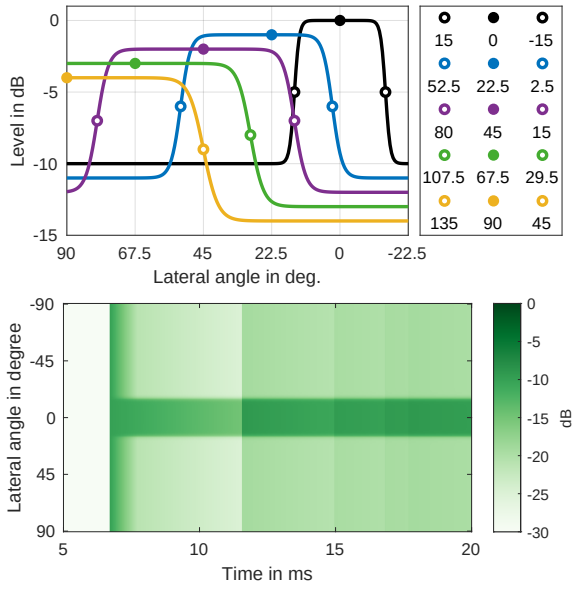

Fig. 2: Masking threshold function for concurrent early reflections. Top: Angular dependency of the threshold function with a dynamic range of 10 dB for direct sound angles of $\phi = \{0°, 22.5°, 45°, 67.5°, 90°\}$ (solid dots). The lateral spread is given by the angle between the open circles (curves are offset in level for visibility). Bottom: Angular-temporal evolution of the threshold function for the case shown in Fig. 3, left.

reflections up to the mixing time—including the floor reflection, which was deemed to be perceptually important [6]—and only a few after the mixing time.

This was achieved with an initial threshold level of $a_0 - 10$ dB, a decay rate of 1 dB/ms starting at the time τ_0 and lateral angle ϕ_0. Similar to the initial grouping of reflections in the first step, a lateral dependency was introduced to mimic a reduced threshold for reflections that are spatially separated from the first sound. Such spatial dependence can be observed in Bech [6, 7] and Begault et al. [19], albeit with some differences regarding the exact nature of the dependence. The angular width of the spatial dependency was again taken from Best et al. [17], while the depth was set to 10 dB [6, 19] (cf. Fig. 2, top). The transition between high and low amplitude threshold was realized with a hyperbolic tangent window. In addition to the lateral dependency, previous studies found that reverberant energy decreases the decay rate of the threshold over time [18, 8]. We chose to model v-shaped thresholds reported by Olive and Toole [18] that exhibit an initial steep negative slope that transitions to a positive slope in the presence of reverberant energy. This was done by adding 35% of the reverberant energy to the threshold, calculated as the RMS energy of all reflections up to the current point in time. The double-sloped threshold suggested by Jensen and Welti [8] was not modeled because it leads to the implausible detection of distinct reflections in late parts of the SRIR as being audible (cf. Fig. 10 in [8]).

Apart from the masking threshold for detecting audible reflections, the proposed model contains an additional echo threshold. In case an echo is detected, the masking threshold resets to the echo's TOA, amplitude, and DOA. The echo threshold is parameterized with a decay rate of 0.06 dB/ms, a depth of 0 dB, without any lateral dependency, and by adding 10% of the reverberant energy to the threshold. This was done to assure that now echos are detected in the SRIR test set. However, a formal evaluation of the echo threshold model is beyond the scope of the current study.

Examples of detected reflections in the empty shoebox rooms (V = 200 m^3, RT = 0.5 s) are shown in Fig. 3. In the case of the *ISM* simulation, the detected first reflections correspond one-to-one to image sources. For $t \gtrsim 30$ ms, multiple SRIR samples are often grouped as one reflection. The temporal dependence of the threshold function, which exhibits the aforementioned v-shape, is visible in the top row of Fig. 3. The lateral dependence is best observed in the center row, where relatively loud contributions around $\phi = 0$ are discarded for $t \gtrsim 15$ ms. While the floor reflection was determined to be audible for all rooms except for the large dry room (RT = 0.5 s, V = 5000 m^3), the ceiling reflection was discarded for all rooms due to the lateral dependence of the threshold function. For the early part of the SRIS the *ISM* and *TRI* simulations appear to be in good agreement. However, spatial smearing of the DOA can be observed in the *TRI* simulations for $t \gtrsim 20 - 25$ ms. This may be caused by the 2 kHz low-pass that was applied to the estimated velocity v (cf. Section 2.1). The low-pass inherently limits the spatio-temporal DOA resolution—i.e., the maximally achievable reflection density—to 2 reflections/ms. For the 200 m^3 room, this limit is reached at 28 ms [13, Eq. 4.5].

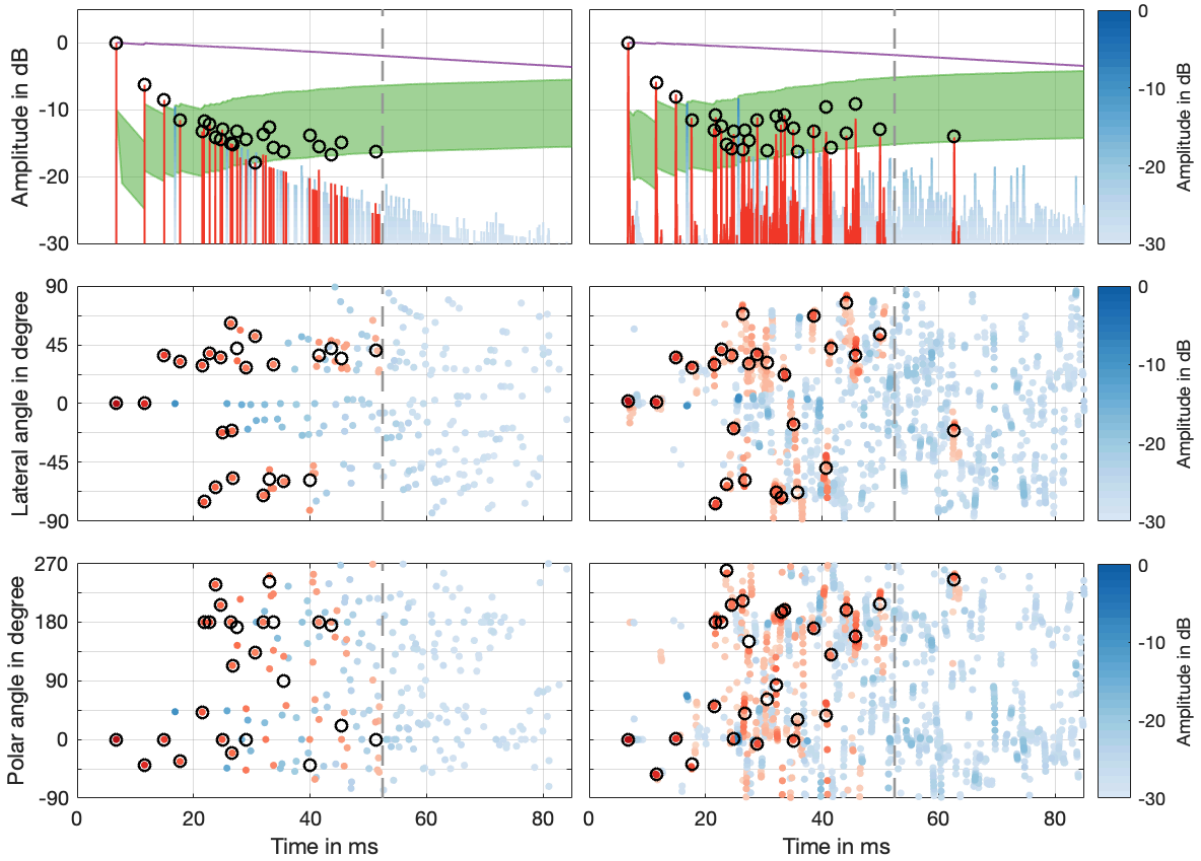

Fig. 3: SRIRs of the small, dry room (RT = 0.5 s, V = 200 m^3) and detected early reflections. Left: *ISM*; Right: *TRI*. SRIRs without spatial information are shown in the top row; the middle and bottom row show the lateral and polar angle, respectively. Potentially audible SRIR contributions are highlighted in red. The TOA, amplitude, and DOA of detected reflections are indicated by black circles. The top row additionally shows the masking threshold function in green and the echo threshold function in purple. The gray dashed line gives the perceptual mixing time.

2.3.3 Selection of early reflections

In the third step, a fixed number of reflections is selected from the list that was extracted in the previous step to account for the available computational resources or desired degree of realism. Three simple selection methods were tested: (i) Use the N *first* reflections, (ii) the N *loudest*, or (iii) the N reflections that *exceed* the masking threshold function the most (cf. Fig. 4). The *first* approach has a tendency to favour early second-order reflections over louder but later-arriving first-order reflections. In the test case this also caused an imbalanced pick of reflections arriving from the left and right. The *exceed* method led to a more balanced selection with respect to the lateral angle, but always discarded the floor reflection. The floor reflection only slightly exceeds the masking threshold function because it arrives soon after the direct sound, which does not give the threshold time to decay. Picking the *loudest* reflections avoided these problems and is similar to Coleman et al. [1, 2].

2.4 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

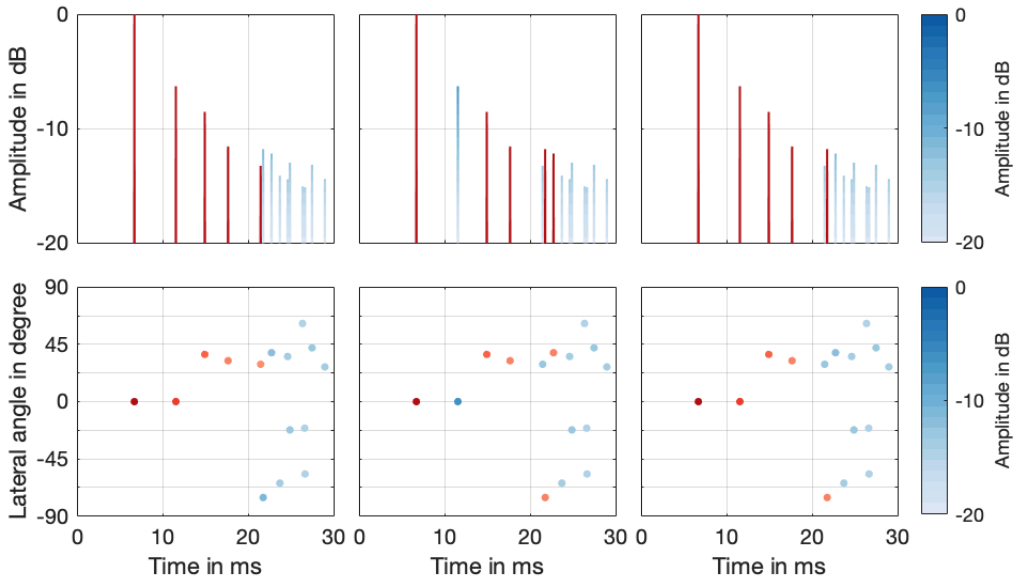

Fig. 4: Selection of $N = 4$ reflections (red) from the list of all initially extracted reflections (blue) for the example of the small dry room and the selection methods *first* (left), *exceed* (middle), and *loudest* (right).

blocks of 256 samples at a sampling rate of 44.1 kHz. For simplicity, the RMS estimate in decibels was approximated by fitting first-order polynomials in a least-squares sense. Two different approaches were considered: Using a single polynomial starting at the position of the last early reflection, and using an additional polynomial starting at the position of the direct sound to account for the residual energy between the direct sound and the last early reflection. In the latter case, the intercept of the additional polynomial was set in a way to assure equal energy at the intersection point, i.e., the position of the latest early reflection (cf. Fig. 5, top).

2.5 Decoding

The direct sound and early reflections were rendered using head-related transfer functions (HRTFs) from the FABIAN HRTF database [20, 21] that were interpolated to the exact DOAs using spherical harmonics of order 35 and added to the binaural room impulse response (BRIR) at τ_i with an amplitude of a_i. For simplicity, the late reverberation was modeled by Gaussian white noise with a diffuse-field interaural cross-correlation [22]. The noise was multiplied with the polynomials estimated from the RMS residual energy to achieve the desired decay function (cf. Fig. 5, bottom). Computationally cheaper possibilities would be to use feedback delay networks [23] or velvet noise [24]. 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].

3 Perceptual evaluation

The performance of the proposed algorithm was evaluated in a listening test consisting of two parts: (i) A study of overall differences across the nine *ISM* rooms, and (ii) a detailed qualitative analysis in one selected room. In all cases, the parametric renderings were directly compared to a reference obtained by a direct binaural rendering of the *ISM* rooms. This was done by applying an HRTF to all simulated image sources together with the *ISM* binaural late reverberation (cf. Section 2.1). This ensured that differences between the test conditions and the reference are likely to become audible due to the direct comparison and can almost exclusively be ascribed to the rendering of early reflections. Examples for auralizations are available online (cf. Section 2).

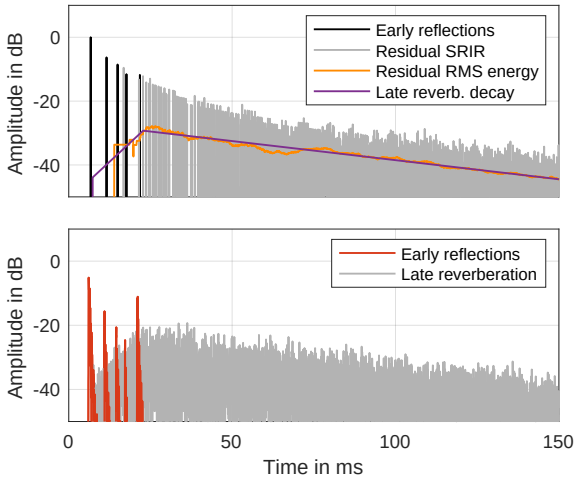

Fig. 5: Parametric rendering of the small dry room using the four *loudest* early reflections and double-sloped late reverberation. Top: SRIR and parametric representation. Bottom: BRIR.

3.1 Listening test stimuli

Parametric renderings with $N = \{0, 1, 6\}$ reflections were chosen as test conditions. An additional parametric condition was generated by manually selecting the direct sound and six first-order reflections (6_{ISM}), using the same decoding as the other test cases (cf. Section 2.5). To limit the duration of the listening test, only the *loudest* method for selecting reflections and the *double-sloped* late reverberation were included as these gave slightly better results than the other methods during informal listening by the authors. As loudness is usually controlled by the user in parametric spatial audio applications, it is irrelevant for the evaluation and was excluded as a cue by equalizing the RMS levels across all test conditions. The gain corrections with respect to the reference were 0.4 dB on average and never exceeded 1.3 dB, which suggests that the encoding preserves the overall SRIR level well. Anechoic male speech (first 5 s from track 50 of the EBU SQUAM CD [2]) was used as the only audio content to limit the duration of the testing phase. Informal listening by the authors suggested that similar results are to be expected for castanets, drums, and string instruments, while differences are likely to be larger for noise signals.

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

3.2 Perceptual metrics

Ten qualities from the Spatial Audio Quality Inventory [25] were selected according to their relevance and completeness based on informal listening by the authors. The qualities are listed in Fig. 6. Detailed descriptions are available from Lindau et al. [25]. The qualities *Difference* and *(difference in) Reverberation* were rated between 0 and 1 with the labels none and large. All other qualities were rated between -1 and 1 with lables shown in Fig. 6.

3.3 Study protocol

Twenty-seven subjects participated in the listening study (5 female, 22 male, average age 38, 18 participated in listening tests before, average of 2.6 hours of audio related tasks per day). The test was conducted in the sound insulated anechoic chamber of Microsoft Research Redmond. It started with an introduction of the ten qualities given by their written circumscriptions [25] during which the subjects could ask questions to clarify their meaning. Next, a brief training was given to familiarize the subjects with rating procedure. The training contained two exemplary stimuli and the corresponding references to cover the range of differences to be expected during the test: $N = 0$ for the large dry room and $N = 6$ for the small wet room. Before the test started, subjects were instructed to always focus on the current quality displayed on the rating interface, take their time at will, listen to the stimuli as often and in any order they wanted, and ask questions if anything was unclear during the test. After the training, the subjects first rated the overall difference between the test conditions and the reference in the nine rooms. The presentation order of the rooms was randomized, and all test conditions for one room ($N = \{0, 1, 6, 6_{ISM}\}$) were presented on a rating screen in randomized order. In the second part the remaining nine qualities were rated for the medium dry room in randomized order. All subjects completed the test within 45 minutes.

3.4 Analysis and results

Fig. 6 shows the median ratings and 95% bootstrapped confidence intervals (non-parametric resampling, bias-corrected and accelerated calculation). Since the test conditions were rated with respect to the reference, a zero-rating denotes no difference and a rating of ± 1 denotes very large differences. Due to the ratings not being normally distributed, multilevel models

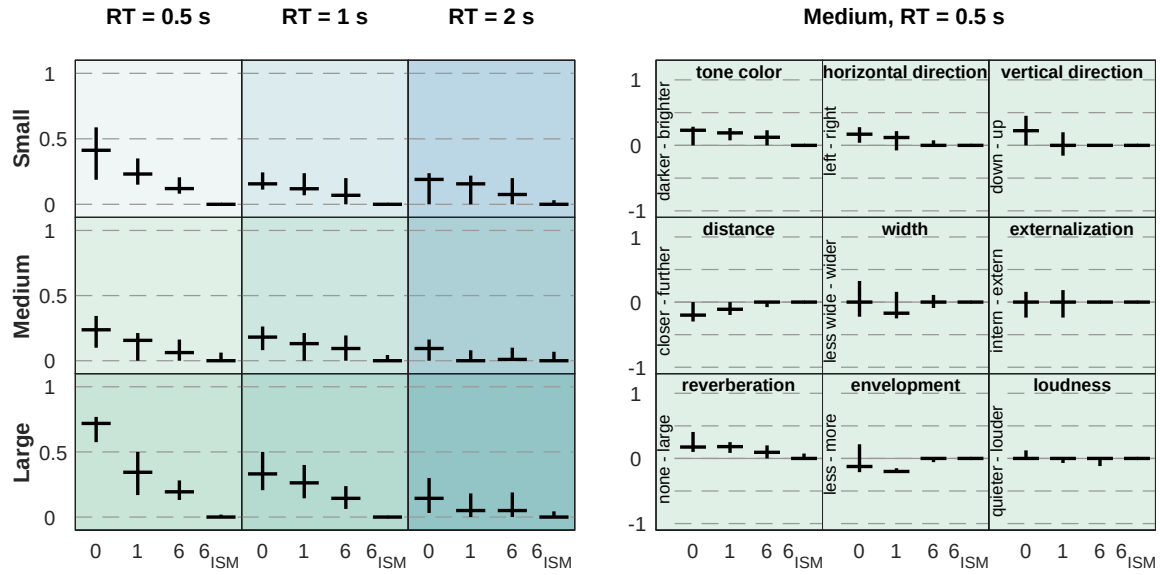

Fig. 6: Results from the listening test given by the median and 95% bootstrapped confidence intervals. Left: *Difference* ratings for all nine rooms. Right: Detailed qualitative ratings for medium, dry room. The qualities are given at the top, the labels on the side. All scales were bi-polar with end points at ± 1 except for *reverberation*, which had endpoints at 0 and 1.

were used for the statistical analysis, which only require normally distributed residuals [26]. The model for *difference* accounts for $R^2 = 49\%$ of the variance (marginal $R^2 = 31\%$ [27]) and the main effects of the three factors *(number of) reflections*, *reverberation*, and *room size* 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 $\hat{\mu} = \{0.31, 0.22, 0.13, 0.05\}$ for $N = \{0, 1, 6, 6_{ISM}\}$), which accounts for 44% of the variance [28, Eq. 20.31]. Notably, the bootstrapped confidence intervals for $N = 6_{ISM}$ overlap with zero in all cases. Differences also decrease with increasing *reverberation* ($\hat{\mu} = \{0.23, 0.17, 0.13\}$ for RT $= \{0.5, 1, 2\}$) accounting for 16% of the variance, whereas the *room size* does not have a clear effect ($\hat{\mu} = \{0.17, 0.14, 0.23\}$ for RT $= \{0.5, 1, 2\}$) accounting for 13% of the variance. Dunn-Šidák corrected pairwise comparisons showed statistically significant differences between all levels for *reflections* and *reverberation*, and between the large room compared to the small and medium room. In addition, all first-order interactions were found to be statistically significant as well. However, interactions were ordinal and thus do not affect the main effects.

The general trend of decreasing differences with increasing number of reflections can also be observed for the detailed analysis conducted only for the medium dry room (cf. Fig. 6, right). In these cases, bootstrapped confidence intervals for $N = 6_{ISM}$ and $N = 6$ always overlap with zero. Multilevel models showed significant effects for the *horizontal direction*, *vertical direction*, *reverberation*, and *loudness*. In the latter case, however, the estimated marginal means differed by only about 5% of the rating scale, which might be considered negligible. Pairwise comparisons showed statistically significant differences for 0 vs. 6_{ISM} and 1 vs. 6_{ISM} but not for 6 vs. 6_{ISM}.

4 Discussion and perspectives

We proposed a parametric encoding of SRIRs with a focus on detecting perceptually-salient early reflections. The proposed encoding can be applied to any SRIR including DOA information, and was initially evaluated against reference simulations obtained from a combined model using image sources and stochastic late reverberation (*ISM*) and a wave based simulation (*TRI*).

The evaluation of the nine simulated empty shoebox rooms showed that the six first-order reflections

$N = 6_{ISM}$ are sufficient for the rendering to be indiscernible from the reference for male speech[3]. This is a reduction of $85-95\%$ considering the $40-118$ image sources before the perceptual mixing time [12]. While the proposed algorithm comes close to the reference if rendering $N = 6$ reflections, the encoding stage needs to be tuned to select a better set of reflections and to be more consistent across rooms. An analysis of the overlap between $N = 6_{ISM}$ and $N = 6$ for each room showed that between 3 and 6 (4.8 on average) first-order reflections were detected initially, but only 2 to 5 (3.4 avg.) were selected using the *loudest* method. Specifically, the ceiling reflection was selected only once in the small, dry room, but never detected in the remaining cases. Reflections from the front and rear wall (as viewed from the receiver, cf. Fig. 1) were detected in most cases but not included in the parametric BRIR in five cases. Reflections from the left wall were not detected in two cases, while reflections from the right wall were always detected but not selected in two cases. The floor reflection was not detected in the large, dry room.

Between 1 and 4 (2.6 avg.) first order reflections were not detected or selected by the *loudest* method. Instead, 1 to 2 (1.1 avg.) prominent second-order reflection and 0 to 3 (1.4 avg.) rather late reflections that aggregated energy across multiple image sources were selected. This might be related to our simplified assumption that sources with the same lateral angle but different polar angles can not be discerned [17], an assumption that has been disproved more recently [30]. Introducing an additional dependency of the masking threshold function shown in Fig. 2 on the polar angle will decrease the spatio-temporal window for detecting reflections. Accordingly, the energy of aggregated reflections will decrease to favor the selection of early reflections over late aggregated reflections. Additionally, it might also help to detect the ceiling reflection that has a spatial separation from the direct sound of about $\Delta\theta = 60^\circ$ for the test cases, which is sufficient to perceive two separate auditory events [30]. Since it might be possible to obtain results that are comparable to $N = 6_{ISM}$ if selecting strong second-order reflections, e.g., for a source or receiver close to a room corner, missing first-order reflections might be less critical in such cases.

[3]Strictly spoken, alternative forced-choice tests are needed to prove inaudibility of differences [29]. Here, we use a less strict definition by assuming indiscernibility if the confidence intervals overlap with 0.

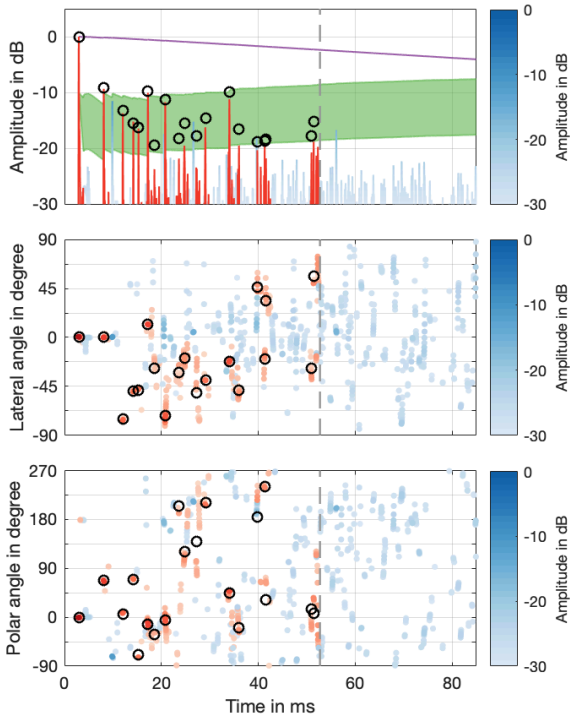

Fig. 7: SRIRs of the meeting room (RT = 0.86 s, V = 224 m^3, Fig. 1, bottom) and detected early reflections (see Fig. 3 for visualization details).

So far, the discussion was restricted to empty shoebox rooms (*ISM*). A comparison of the small empty shoebox simulations *ISM* and *TRI* rooms across the three reverberation times showed that the reflections detected by the proposed encoding scheme were almost identical up to approximately 20 ms, which is close to the time limit for spatio-temporal smearing of the DOA estimates (cf. Section 2.3.2). After 20 ms, the detected reflections diverge and larger levels are found for the *TRI* rooms because more SRIR samples fall into the spatio-temporal window used for level estimation due to the smearing. Accordingly, the *loudest* method has a tendency to select these reflections instead of earlier ones, and only the first-order floor and left-wall reflections are included in the parametric *TRI* BRIRs. An initial inspection of reflections detected in the meeting room suggests that the encoding also works in non-empty rooms but suffers from the smearing as well (cf. Fig. 7). This problem might be at least partially solved by introducing the additional dependency of the threshold function on the polar angle (see above),

and a stricter time limit for detecting early reflections—aspects that deserve more attention in future studies.

Apart from including frequency-dependent rendering [2, 3] and directional late reverberation [4] in the encoding, it might be interesting to consider the spatial variance of the DOA $\sphericalangle_i$ in cases where the inherent point-source assumption does not hold, i.e., when a reflection is spatially widened due to scattering or diffraction. Additionally, cases where the echo threshold function becomes relevant should be included in future evaluations. In contrast to Coleman et al. [2], our algorithm is able to detect multiple reflections in close temporal proximity due to spatial windowing (cf. Fig. 2) but was not yet tested on measured SRIRs. Considering absolute sound levels as Green and Kahle [31] will be difficult assuming that the playback level will be user-controlled in most cases.

Incorporating the proposed approach in virtual or mixed reality (MR) applications requires further work. Head rotations are challenging, because the spatial width of the threshold function (cf. Fig. 2, top) depends on the DOA that is calculated with respect to the listener orientation. This might be solved by moving the spatial dependency from the encoding to the decoding stage, which, as a side effect, will cause more detected reflections. For scenarios where the source and listener may translate as well, parameters must be encoded offline for numerous spatial source-listener location pairs in large scenes [4]. This poses additional challenges regarding spatial smoothness and sensitivity to band limitation. Initial investigations in the meeting room suggest 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. Such jumps in encoded representation hurt spatial compression and runtime memory use. It will also have to be tested how well our proposed encoding works for wave-based simulations on larger scenes that are typically band-limited to 1 kHz [4] rather than 8 kHz employed here, which results in substantially reduced temporal resolution.

References

[1] Coleman, P., Franck, A., Jackson, P. J. B., Hughes, R. J., Remaggi, L., and Melchior, F., "Object-Based Reverberation for Spatial Audio," *J. Audio Eng. Soc*, 65(1/2), pp. 66–77, 2017, doi:10.17743/jaes.2016.0059.

[2] Coleman, P., Franck, A., Menzies, D., and Jackson, P. J. B., "Object-based reverberation encoding from first order Ambisonics RIRs," in *142nd AES Convention, Convention Paper 9731*, Berlin, Germany, 2017.

[3] Stade, P., Arend, J., and Pörschmann, C., "Perceptual evaluation of synthetic early binaural room impulse responses based on a parametric model," in *142nd AES Convention, Convention Paper 9688*, Berlin, Germany, 2017.

[4] Raghuvanshi, N. and Snyder, J., "Parametric directional coding for precomputed sound propagation," *ACM Trans. Graph.*, 37(4), p. Article 108, 2018, doi:10.1145/3197517.3201339.

[5] Godin, K. W., Gamper, H., and Raghuvanshi, N., "Aesthetic modification of room impulse responses for interactive auralization," in *AES Conf. on Immersive and Interactive Audio*, p. Conference Paper 44, York, UK, 2019.

[6] Bech, S., "Timbral aspects of reproduced sound in small rooms. I," *J. Acoust. Soc. Am.*, 97(3), pp. 1717–1726, 1995, doi:10.1121/1.413047.

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

[8] Jensen, R. E. and Welti, T. S., "The importance of reflections in a binaural room impulse response," in *114th AES Convention*, p. Convention Paper 5839, Amsterdam, The Netherlands, 2003.

[9] Buchholz, J. M., Mourjopoulos, J., and Blauert, J., "Room masking: Understanding and modelling the masking of room reflections," in *110th AES Convention, Convention Paper 5312*, Amsterdam, The Netherlands, 2001.

[10] Röhrbein, M. and Lindau, A., "Reducing the temporal resolution of spatial impulse responses with an auditory model," in *Fortschritte der Akustik – DAGA 2011*, pp. 327–328, Düsseldorf, Germany, 2011.

[11] Brinkmann, F. and Weinzierl, S., "AKtools – An Open Software Toolbox for Signal Acquisition, Processing, and Inspection in Acoustics," in *142nd AES Convention, Convention e-Brief 309*, Berlin, Germany, 2017.

[12] Lindau, A., Kosanke, L., and Weinzierl, S., "Perceptual evaluation of model- and signal-based predictors of the mixing time in binaural room impulse responses," *J. Audio Eng. Soc.*, 60(11), pp. 887–898, 2012.

[13] Kuttruff, H., *Room acoustics*, Spon Press, Oxford, UK, 5th edition edition, 2009.

[14] Blauert, J., *Spatial Hearing. The psychophysics of human sound localization*, MIT Press, Cambridge, Massachusetts, revised edition, 1997.

[15] Ziegelwanger, H. and Majdak, P., "Modeling the direction-continuous time-of-arrival in head-related transfer functions," *J. Acoust. Soc. Am.*, 135(3), pp. 1278–1293, 2014.

[16] Ege, R., van Opstal, A. J., Bremen, P., and van Wanrooij, M. M., "Testing the precedence effect in the median plane reveals backward spatial masking of sound," *Scientific Reports*, 8, p. Article 8670, 2018, doi:10.1038/s41598-018-26834-2.

[17] Best, V., van Schaik, A., and Carlile, S., "Separation of concurrent broadband sound sources by human listeners," *J. Acoust. Soc. Am.*, 115(1), pp. 324–336, 2004, doi:10.1121/1.1632484.

[18] Olive, S. E. and Toole, F. E., "The detection of reflections in typical rooms," *J. Audio Eng. Soc.*, 37(7/8), pp. 539–553, 1989.

[19] Begault, D. R., McClain, B. U., and Anderson, M. R., "Early reflection thresholds for anechoic and reverberant stimuli within a 3-D sound display," in *18th Int. Congress on Acoustics (ICA)*, Kyoto, Japan, 2004.

[20] Brinkmann, F., Lindau, A., Weinzierl, S., Par, S. v. d., Müller-Trapet, M., Opdam, R., and Vorländer, M., "A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations," *J. Audio Eng. Soc*, 65(10), pp. 841–848, 2017, doi:10.17743/jaes.2017.0033.

[21] Brinkmann, F., Lindau, A., Weinzierl, S., Geissler, G., van de Par, S., Müller-Trapet, M., Opdam, R., and Vorländer, M., "The FABIAN head-related transfer function data base," DOI: 10.14279/depositonce-5718.3, 2017.

[22] Borß, C. and Martin, R., "An improved parametric Model for perception-based design of virtual acoustics," in *AES 35th International Conference*, London, UK, 2009.

[23] Steffens, H., van de Par, S., and Ewert, S. D., "Perceptual relevance of speaker directivity modelling in virtual rooms," in *Int. Congress on Acoustics (ICA)*, pp. 2651–2658, Aachen, Germany, 2019.

[24] Välimäki, V., Holm-Rasmussen, B., Alary, B., and Lehtonen, H.-M., "Late reverberation synthesis using filtered velvet noise," *Applied Sciences*, 7(5), p. 483, 2017, doi:http://doi.org/10.3390/app7050483.

[25] Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkmann, F., and Weinzierl, S., "A Spatial Audio Quality Inventory (SAQI)," *Acta Acust. united Ac.*, 100(5), pp. 984–994, 2014, doi:10.3813/AAA.918778.

[26] Hox, J. J., *Multilevel Analysis. Techniques and Apllications*, Quantitative Methodology, Routledge, New York, Hove, 2 edition, 2010.

[27] Nakagawa, S. and Schielzeth, H., "A general and simple method for obtaining R^2 from generalized linear mixed-effects models," *Methods in Ecology and Evolution*, 4, pp. 133–142, 2013, doi:https://doi.org/10.1111/j.2041-210x.2012.00261.x.

[28] Eid, M., Gollwitzer, M., and Schmitt, M., *Statistik und Forschungsmethoden*, Beltz, Basel, Switzerland, 5 edition, 2017.

[29] Leventhal, L., "Type 1 and type 2 errors in the statistical analysis of listening tests," *J. Audio Eng. Soc.*, 34(6), pp. 437–453, 1986.

[30] Pulkki, V., Pöntynen, H., and Santala, O., "Spatial Perception of Sound Source Distribution in the Median Plane," *J. Audio Eng. Soc*, 67(11), pp. 855–870, 2019, doi:10.17743/jaes.2019.0033.

[31] Green, E. and Kahle, E., "Dynamic spatial responsiveness in concert halls," *Acoustics*, 1(3), pp. 549–560, 2019, doi:10.3390/acoustics1030031.