CHALMERS
UNIVERSITY OF TECHNOLOGY

Perceptual Evaluation of Spatial Room Impulse Response Extrapolation by Direct and Residual Subspace Decomposition

Audio Engineering Society

Conference Paper

Presented at the 2022 International Conference on
Audio for Virtual and Augmented Reality
2022 August 15–17, Redmond, WA, USA

Perceptual Evaluation of Spatial Room Impulse Response Extrapolation by Direct and Residual Subspace Decomposition

Thomas Deppisch[1], Sebastià V. Amengual Garí[2], Paul Calamia[2], and Jens Ahrens[1]

[1]*Chalmers University of Technology, 412 96 Gothenburg, Sweden*
[2]*Reality Labs Research, Meta, Redmond, WA 98052, USA*

Correspondence should be addressed to Thomas Deppisch (thomas.deppisch@chalmers.se)

ABSTRACT

Six-degrees-of-freedom rendering of an acoustic environment can be achieved by interpolating a set of measured spatial room impulse responses (SRIRs). However, the involved measurement effort and computational expense are high. This work compares novel ways of extrapolating a single measured SRIR to a target position. The novel extrapolation techniques are based on a recently proposed subspace method that decomposes SRIRs into a direct part, comprising direct sound and salient reflections, and a residual. We evaluate extrapolations between different positions in a shoebox-shaped room in a multi-stimulus comparison test. Extrapolation using a residual SRIR and salient reflections that match the reflections at the target position is rated as perceptually most similar to the measured reference.

1 Introduction

Six-degrees-of-freedom (6DoF) rendering of an acoustic environment allows a listener to freely move through the environment and rotate their head. The renderer thus aims at reproducing the acoustic properties of the environment at arbitrary locations and for arbitrary head orientations. In extended reality (XR) applications, such renderers are often required to reproduce audio signals in an acoustic environment that is different from the one they were recorded in. The linear, time-invariant characteristics of an acoustic source-receiver transfer path in any environment can be imposed on a signal by convolution with the respective impulse response. As model-based methods for the calculation of room impulse responses (RIRs) are usually not able to create a perceptually authentic reproduction of acoustic environments [1], typically having difficulties with reproducing the original spectrum and spatial properties, renderers often utilize measured RIRs to reproduce the acoustic properties of a real-world environment. Most often, the renderers target binaural headphone playback and thus involve the convolution of signals with binaural room impulse responses (BRIRs).

The direct measurement and interpolation of BRIRs

from a dense grid of positions and head orientations involves an impractically high measurement effort. A recent study comparing different BRIR modifications however suggests that parametric modification of a reduced set of BRIRs may allow for a perceptually plausible rendering [2]. Specifically, the reproduction of the direct-to-reverberant energy ratio (DRR) and of a sufficient amount of reverberant energy was found to be essential, while the pre-delay of the BRIR, the initial time-delay gap (ITDG) and the spatio-temporal characteristics of reflection patterns turned out to be less important. A plausible BRIR-based rendering might hence be achieved by combining a synthesized or interpolated direct sound with a single, static BRIR tail and modifying the energy ratio between the two to adjust the DRR. This approach delivered promising results in terms of plausibility [3] and also transfer-plausibility [4], i.e., participants showed low detection rates when asked to distinguish between *real* sound sources that were played back via loudspeakers and *synthetic* sound sources that were rendered via headphones. However, a similar experiment [5] that allowed the participants to freely move within a listening area found lower transfer-plausibility in most cases.

As the directional information of measured BRIRs is encoded into the spectral characteristics of head-related transfer functions (HRTFs), accessing and parametrically modifying that information is difficult. Furthermore, as measured BRIRs inherently involve a specific set of HRTFs, direct BRIR-based rendering does not allow for the user-dependent switching of HRTFs. Instead of modifying measured BRIRs, the method from [6] proposes synthesizing BRIRs from a single omnidirectional RIR. The authors report promising results from a numerical evaluation but limited plausibility in a perceptual pilot study. As neither directional nor spatial information is captured, the method heavily relies on a geometric model.

A more flexible synthesis of BRIRs in combination with lower measurement effort can be achieved by employing compact microphone arrays, i.e., microphone arrays with small aperture that facilitate the simultaneous measurement of a set of RIRs to capture the directional properties of the acoustic transfer path at a single receiver position. Most commonly used are spherical microphone arrays (SMAs) and the set of captured RIRs is often referred to as spatial room impulse response (SRIR). With suitable processing, SRIRs allow for the synthesis of BRIRs for arbitrary head rotations. Accurate and HRTF-flexible 6DoF rendering can then be achieved by the interpolation of SRIRs from a dense grid of measurement positions [7, 8, 9]. Some of the referenced methods attempt to reduce the resulting measurement and computational effort by only interpolating early reflections and pre-interpolating a set of measurement positions to a denser set.

The present study investigates the perceptual quality of different methods that attempt to perform an acoustic translation from one position in a room to a target position, based on a single SRIR measurement. As the directional acoustic room response is only available at a single position, a renderer based on this premise requires additional information that could be obtained from a geometric model or via parameter estimation from the measurement. In this contribution, we attempt to lay the groundwork for such a renderer but disregard the problem of parameter estimation. Instead, we evaluate the perceived quality of different SRIR extrapolations and, depending on the method, assume that some information at the target position is given. This given information includes the spatio-temporal properties of the direct sound, the DRR, and, for one of the extrapolation methods, the spatio-temporal properties of salient reflections. We use the term salient reflections to refer to reflections that energetically stand out against the superposition of weaker reflections at any given time in an RIR.

Some of the methods under investigation exploit the individual access to the direct part and the residual of the SRIR that are obtained via the recently proposed direct and residual subspace decomposition method for SRIRs [10]. The direct part SRIR comprises the direct sound and salient reflections, while the residual SRIR comprises what is left after the direct part has been removed from the SRIR. SRIR extrapolation may be based on the assumption that the residual SRIR stays constant throughout the room, while the direct sound and salient reflection patterns change significantly.

To evaluate the different extrapolation methods, we asked participants to rate the perceived similarity of the methods compared to a measured reference in a multi-stimulus comparison test. The extrapolated SRIRs were pre-rendered for multiple static positions and rendered with dynamic, three-degrees-of-freedom (3DoF) head tracking. We hence refer to the process as SRIR extrapolation rather than 6DoF rendering. The results suggest that the addition of salient reflections that match the reference is a significant improvement over rendering with

unmatched or without salient reflections. In addition, SRIRs with unmatched and without salient reflections are rated significantly closer to the reference than a static BRIR rendering with matched direct sound in a majority of the tested conditions.

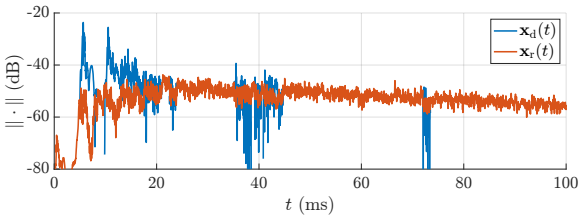
2 Direct and Residual Subspace Decomposition of SRIRs

The authors recently proposed the direct and residual subspace decomposition for SRIRs [10]. The method decomposes an SRIR into a direct part that comprises the direct sound and salient reflections, and a residual. The residual typically contains a temporally increasing amount of non-salient reflections, non-transient components of the room response, e.g. due to room modes, and noise. Both the direct part and the residual are obtained as SRIRs with the same number of channels as the original SRIR and the decomposition preserves their interchannel relationships. Compared to another recently proposed method [11] that utilizes a beamformer and exploits a comprehensive signal model in the spherical harmonic (SH) domain, the direct and residual subspace decomposition provides an increased separation performance as it is not relying on accurate direction-of-arrival (DOA) estimates for reflections.

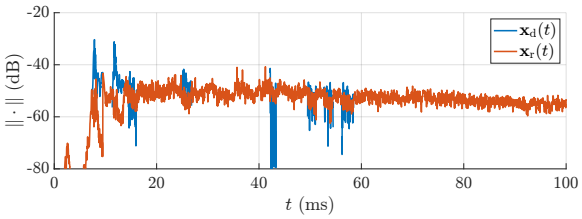
In a nutshell, the decomposition is performed by a generalized singular value decomposition (GSVD) of the SRIR and a current estimate of the residual. The decomposition is started at the end of the SRIR, where a first estimate of the residual is taken, and then applied in a blockwise manner while proceeding toward the beginning of the SRIR and updating the estimate. The sum of the direct part $\boldsymbol{x}_\mathrm{d}(t)$ and the residual $\boldsymbol{x}_\mathrm{r}(t)$ perfectly reconstructs the original SRIR, $\boldsymbol{x}(t) = \boldsymbol{x}_\mathrm{d}(t) + \boldsymbol{x}_\mathrm{r}(t)$. Fig. 1 shows the norms of the direct part $\boldsymbol{x}_\mathrm{d}(t)$ and the residual $\boldsymbol{x}_\mathrm{r}(t)$ of three SRIRs that are obtained by the subspace decomposition of three 25-channel SRIRs in the SH domain. The SRIRs stem from a public dataset [12] that is described in Sec. 3.1 and is used throughout the listening experiment.

The direct and residual subspace decomposition can either be directly applied to the array signals or to an SH decomposition thereof. In this contribution, we solely use SH-domain SRIRs to be able to use common SH-domain processing methods, including a rotation of the sound field and binaural rendering.

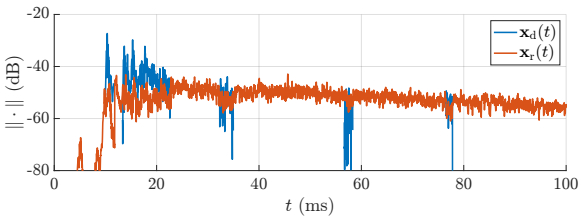
For all SRIRs, the direct and residual subspace decomposition was performed with a block size of 64 samples



(a) Receiver *R2*, source *S1*



(b) Receiver *R2*, source *S2*



(c) Receiver *R1*, source *S3*

Fig. 1: Norms of the direct part $\boldsymbol{x}_\mathrm{d}(t)$, comprising the direct sound and salient reflections, and the residual $\boldsymbol{x}_\mathrm{r}(t)$ of 25-channel SH-domain SRIRs measured for different source and receiver positions in a shoebox-shaped room.

(1.3 ms) and a hop size of 8 samples. The residual estimate had a length of 20 ms and the thresholds for the detection of salient reflections were calculated by averaging the generalized singular values over 32 blocks and using a margin of 3 times their averaged standard deviation. We refer the reader to [10] for a detailed description of those parameters. The decomposition was only applied to the early part of the SRIRs up until 80 ms to reduce the computational effort and the estimated number of subspace components was smoothed over 7 blocks to increase the extraction windows of salient reflections.

A reference implementation[1] of the direct and residual subspace decomposition method and a companion website with listening examples[2] are available.

[1] https://github.com/thomasdeppisch/SRIR-Subspace-Decomposition

[2] http://www.ta.chalmers.se/SRIR-subspace-decomposition/

3 SRIR Extrapolation

3.1 SRIR Dataset

In this contribution, we use SRIRs from a publicly available dataset that were measured in a variable-acoustics room of dimensions $7.87 \times 5.75 \times 2.91$ m [12]. In particular, we use the measurements with an active absorption of 25%, resulting in reverberation times of 0.72 s, 0.76 s and 0.84 s, at 500 Hz, 1 kHz and 2 kHz, respectively. The measurements were performed with Genelec 8331A coaxial loudspeakers and we selected SRIRs that were captured with the Eigenmike em32 microphone array. The SRIRs are provided in the SH domain including up to fourth-order SHs.

The 3 source positions and the 5 receiver positions that are used in this work are illustrated in Fig. 2. Note that the original dataset contains 7 receiver positions and the receivers with numbers *R1* to *R5* do not coincide with the receivers *R1* to *R5* of the original dataset. Additionally, Fig. 2 shows the three extrapolation pairs *E1–E3*, i.e., pairs of receiver positions that will be used in the following to evaluate the extrapolation methods. As illustrated in the figure, the extrapolation *E1* comprises taking the SRIR measured at position *R2* with source *S1* and modifying it to approximate the SRIR at the target position *R5*. Similarly, *E2* is an extrapolation from position *R2* to target position *R4* with source *S2*, and *E3* is an extrapolation from position *R1* to position *R3* with source *S3*. The norms of the direct part SRIRs and residual SRIRs of the three initial positions of the extrapolation pairs *E1–E3* are shown in Fig. 1.

There are qualitative differences between the extrapolation pairs. In the case of the pairs *E2* and *E3*, both corresponding receiver positions are at a similar distance from the source position and, assuming a loudspeaker directivity that is symmetric around the main axis of radiation, are similarly influenced by the loudspeaker directivity. As a result, the DRRs of the initial position and the target position are similar in the case of those two extrapolation pairs. In the case of the extrapolation pair *E1*, the DRRs of the initial position *R2* and the target position *R5* differ considerably, mainly due to the loudspeaker directivity which leads to attenuation of the direct sound at the target position *R5*.

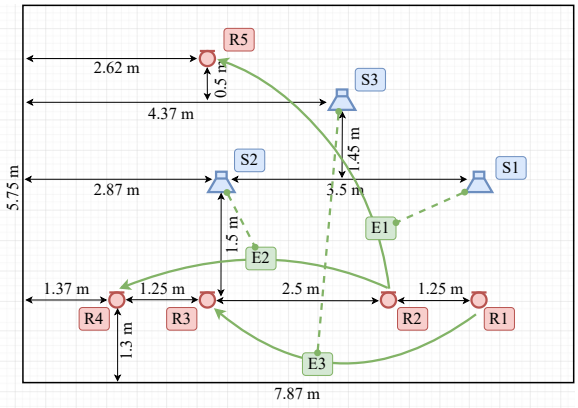


Fig. 2: Source positions *S1–S3* and receiver positions *R1–R5* of the SRIR measurements from [12] as well as the extrapolation pairs *E1–E3*.

3.2 SRIR Extrapolation Methods

In the following, the different extrapolation methods under test are introduced. They will be compared perceptually in the listening experiment in Sec. 4. The extrapolation is always performed by taking an SRIR from one receiver position and modifying it to approximate an SRIR that was measured at another receiver position, which is also referred to as the target position. The SRIR that was measured at the target position thus serves as a reference. The extrapolation methods are summarized in Tab. 1.

All extrapolation methods have in common that they contain the direct sound from the target SRIR. Perceptual differences that are solely attributed to the direct sound are not of interest in this study, as it is assumed that an SRIR-based renderer individually processes the direct sound of a source to be able to model arbitrary source directivities and distance attenuation. What is more, all extrapolation methods modify the DRR to fit the target DRR by scaling the energy of the SRIR in relation to its direct sound. On the other hand, the ITDG does not match the target ITDG in all methods except *trans*, where the direct sound and salient reflections of the target SRIR are employed.

3.2.1 Static BRIR Plus Direct Sound (*stat*)

The extrapolation method *stat* is based on a pre-rendered binaural room impulse response (BRIR) and

Table 1: Summary of the extrapolation methods.

stat	Static BRIR with dynamically rendered direct sound from the target SRIR and matched DRR.
rot	SRIR, rotated to match the direct sound direction with the target SRIR, with the direct sound from the target and matched DRR.
res	Residual SRIR, rotated to match the direct sound direction with the target SRIR, with the direct sound from the target and matched DRR.
trans	Residual SRIR, rotated to match the direct sound direction with the target SRIR, with the direct sound and salient reflections from the target and matched DRR.

serves as a baseline method. A similar method was investigated in terms of (transfer-)plausibility in [3, 4, 5]. The BRIR is obtained by cutting off the direct sound from a measured SRIR, scaling the rest to fit the reverberant energy of the target SRIR and rendering it via the magnitude least squares (magLS) binaural rendering method [13]. The extrapolation to the target position is then achieved by adding a dynamically rendered direct sound from the 25-channel SH-domain target SRIR to the 2-channel BRIR. As the direct sound part of the full SRIR is employed, this part can be freely rotated during rendering, while the 2-channel BRIR cannot and is hence termed static. Due to the scaling of the SRIR prior to creating the pre-rendered BRIR, the DRR of the extrapolation equals the target DRR after the combined rendering.

3.2.2 Rotated SRIR Plus Direct Sound (*rot*)

The extrapolation method *rot* is similar to the first method *stat* but employs the full measured SRIR instead of a pre-rendered BRIR. The extrapolation is performed by rotating the measured SRIR in the SH domain to align its direct sound with the direct sound of the target SRIR, removing the direct sound and replacing it with the direct sound of the target SRIR. The reverberant part is again scaled so that the DRR of the extrapolated SRIR matches the DRR of the target.

3.2.3 Rotated Residual Plus Direct Sound (*res*)

The extrapolation method *res* is based on the direct and residual subspace decomposition from Sec. 2. *Res* is similar to *rot* but instead of the measured SRIR, only the residual SRIR is involved in the processing and thus the rendered response does not contain any salient reflections. The extrapolation comprises rotating the residual SRIR in the SH domain to align the direction where direct sound would have been with the direct sound of the target SRIR, adding the direct sound from the target, and scaling the residual SRIR such that the DRR matches the target DRR. The rotation is based on the assumption that the strongest directionality of the residual appears in the direction of the direct sound. The directionality of anisotropic late reverberation may thus be reproduced in a physically inaccurate direction.

3.2.4 Rotated Residual Plus Salient Reflections And Direct Sound (*trans*)

The extrapolation method *trans* is also based on the direct and residual subspace decomposition and aims at recreating the acoustic translation as accurately as possible by including salient reflections from the target SRIR. The extrapolation is achieved by modifying the measured residual SRIR in the same way as in the *res* extrapolation method but instead of adding only the direct sound from the target SRIR, the full direct part SRIR of the target SRIR, including direct sound and salient reflections, is added to the rotated and scaled residual. This extrapolation method contains the idealistic assumption that the spatio-temporal characteristics of the direct sound and the salient reflections at the target position are known. In practice, a renderer can only estimate those properties. This study thus investigates a best-case scenario.

4 Listening Experiment

4.1 Experiment Design

We evaluate the SRIR extrapolation methods from Sec. 3 using a multi-stimulus comparison test with hidden reference. In contrast to the commonly used multi-stimulus test with hidden reference and anchor (MUSHRA) [14], we chose not to employ an anchor condition that typically is created by low-pass filtering the reference condition and hence would exhibit a fundamentally different type of degradation compared to the other conditions. In a recent comparison of several audio quality evaluation paradigms in virtual reality scenes that included both direct and indirect scaling

methods, such an anchor-less multi-stimulus comparison test revealed the highest number of significant differences between conditions [15].

The similarity of the conditions is rated on a quasi-continuous scale with the labels *identical*, *very similar*, *similar*, *different* and *very different* that allows for ratings between the labels. For the evaluation, the ratings are converted to an integer score between 0 (*very different*) and 100 (*identical*).

4.2 Conditions And Trials

In each trial, the participants rated the similarity of 5 conditions to the reference, where one of the conditions was the hidden reference. The conditions were created by convolution of the target SRIR and the extrapolated SRIRs with either a speech or a drum sample. The speech sample contained male speech and was taken from the Sound Quality Assessment Material recordings (EBU SQAM), while the drum sample contained a drum loop that was extracted from stems of the song Spicy Funk Cake[3]. The combination of 2 stimuli types (*speech* and *drums*) and 3 extrapolation pairs *E1–E3* (cf. Fig. 2) results in a total of 6 trials. For each target position, the sound fields of the reference SRIR and the extrapolated SRIRs were rotated in the SH domain such that the direct sound appeared directly in front of the listener. Additionally, each trial was repeated once with an applied yaw rotation of the sound field by $-60°$, yielding a total number of $2 \cdot 3 \cdot 2 = 12$ trials.

4.3 Setup

The experiment was implemented using an open-source software tool[4]. The software is specifically designed to perform multi-stimulus comparison tests and can be configured using a text file in the JavaScript Object Notation (JSON) format that describes the experiment. Fig. 3 shows the user interface that is automatically generated by the software based on the JSON configuration. The software acts as a remote control sending out Open Sound Control (OSC) messages to a rendering software. We used Reaper[5] to play back the conditions, and the SceneRotator and BinauralDecoder plugins of the IEM

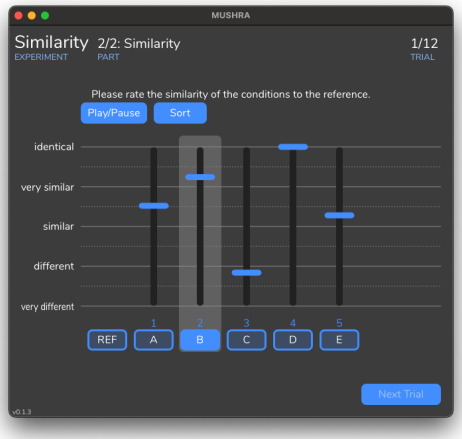


Fig. 3: The user interface allows for the rating of the similarity of 5 conditions to the reference.

Plugin Suite[6] to apply a 3DoF sound field rotation according to head tracker data as well as to perform the binaural rendering of the conditions.

The software was configured to start with a training session comprising two trials, one with the speech-based and one with the drum-based conditions. After the training session, the experiment started and the conditions had to be rated in 12 trials. The order of the conditions in each trial as well as the order of the trials was randomized. The conditions were continuously looped, the participants could seamlessly switch between conditions at any time and they were allowed to take as much time as they needed. A condition could be selected for playback either by a mouse click or by a keyboard shortcut and the sliders could be re-ordered by decreasing rating by clicking the *sort* button.

4.4 Procedure

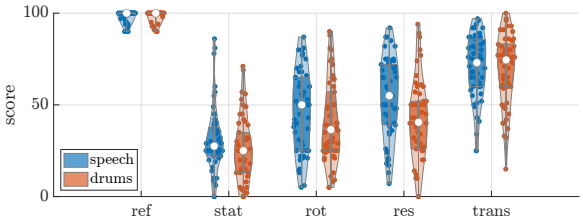
10 participants aged between 23 and 43 years with an average of 31.5 years took part in the experiment. The duration of the experiment was between 14 and 29 minutes, with an average of 21 minutes. 3 participants reported having more than 5 years of experience in listening to binaural audio and are hence considered expert listeners. 3 participants reported having between 3 and 5 years of experience and are considered experienced listeners. All of the participants reported having previously participated in listening experiments more

[3] https://doi.org/10.5281/zenodo.3601032
[4] https://git.iem.at/rudrich/mushra
[5] http://reaper.fm/
[6] https://plugins.iem.at/

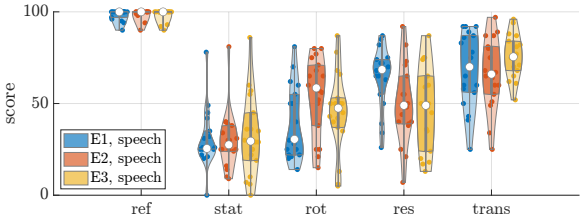
than twice. None of the participants reported a known hearing impairment. One participant rated the hidden reference with a score of less than 90 in more than 15% of the cases and was thus excluded from the results below. After the experiment, the participants were asked to report their strategy to rate the different conditions and what properties of the conditions they concentrated on.

4.5 Results

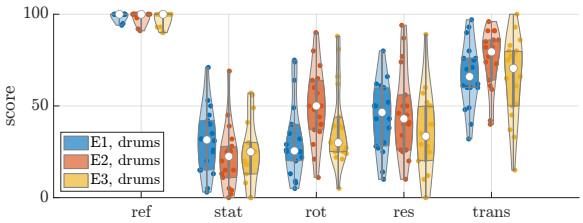
Fig. 4 shows violin plots of the results of the listening experiment for, (a), the *speech* and the *drum* stimuli pooled over all three extrapolation pairs, (b), the individual extrapolation pairs with the *speech* stimulus, and (c), the individual extrapolation pairs with the *drum* stimulus. The violin plots show the median scores as white circles and individual scores as colored circles. The shape of the violins is determined by a kernel density estimate of the underlying distribution. The interquartile ranges between the upper and lower quartiles of the scores are illustrated by opaque boxes and a thick gray line.



(a) Scores of the *speech* and the *drum* stimuli pooled over all extrapolation pairs.



(b) Scores of the extrapolation pairs *E1–E3* for the *speech* stimulus.



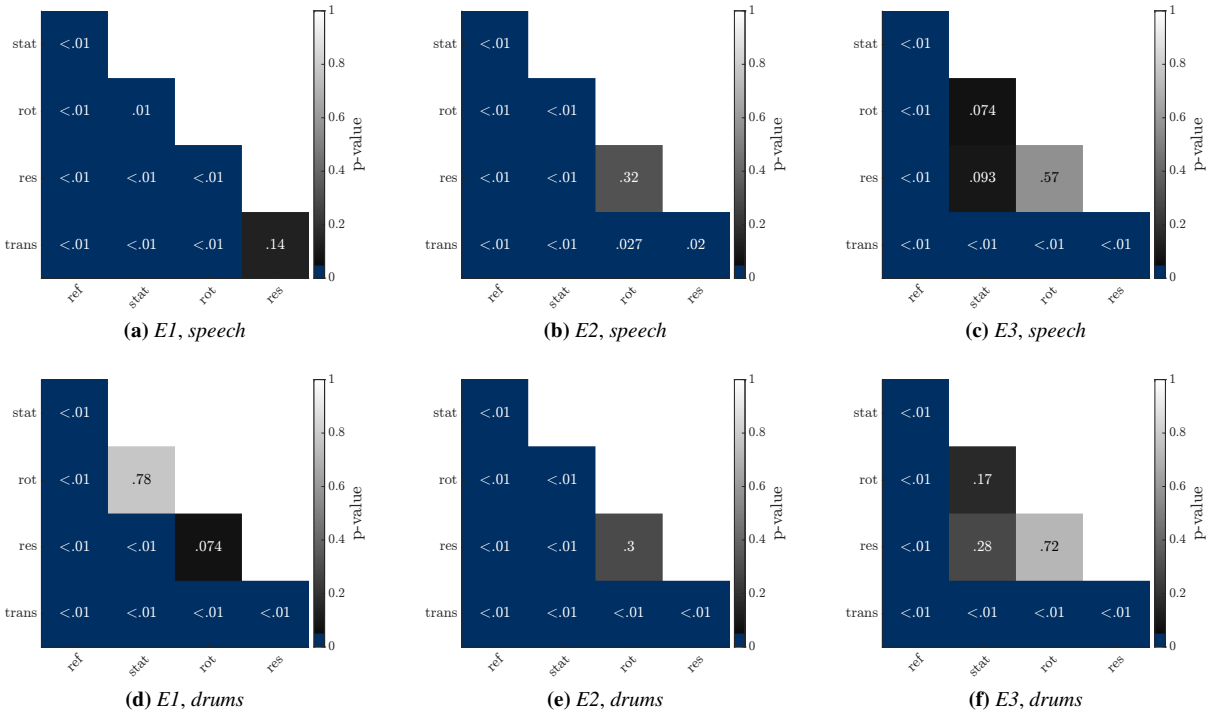
(c) Scores of the extrapolation pairs *E1–E3* for the *drum* stimulus.

Fig. 4: Violin plots of the listening experiment results show the median of the scores as a white circle, individual ratings as colored circles and the interquartile range as opaque box.

In all cases, the hidden reference *ref* reached a median score of 100 and the extrapolation *trans* reached the highest median among the different extrapolation methods. In all but one cases (Fig. 4c, *E1*), the extrapolation *stat* received the lowest median scores. Furthermore, when analyzing the pooled results in Fig. 4a, the extrapolation *res* shows slightly higher median scores than *rot*, but when looking at the results of the individual extrapolation pairs, Figs. 4b and 4c, this is only true for the extrapolation pairs *E1* and *E3*, not for *E2*. The pooled results for *speech* and *drum* stimuli in Fig. 4a show similar distributions for both stimulus types but interestingly, there is a slight trend for a higher median score in case of the *speech*-based conidtions.

To gain further insights into the results, a statistical analysis was performed. The Shapiro-Wilk test for normality suggested non-normally distributed data at a 5% significance level for 33% of the 30 data sets of the extrapolations *E1–E3* that are also shown in Figs. 4b and 4c. 60% of the non-normally distributed data can be attributed to the ratings of the hidden reference, which are typically non-normally distributed due to ceiling effects [16]. Further analysis was therefore performed using the non-parametric Friedman test. The Friedman test found significant differences at $p < 0.01$ for all three extrapolations *E1–E3* for the *speech* stimulus ($\chi^2(4) = 57.1$, $p < 0.01$, $\chi^2(4) = 56.3$, $p < 0.01$ and $\chi^2(4) = 56.24$, $p < 0.01$) and also for the *drum* stimulus ($\chi^2(4) = 60.08$, $p < 0.01$, $\chi^2(4) = 60.02$, $p < 0.01$ and $\chi^2(4) = 51.06$, $p < 0.01$). Thus, pairwise Wilcoxon Signed Rank tests were conducted as post-hoc measure. The resulting p-values were adjusted according to the Bonferroni-Holm p-value correction and are displayed in Fig 5. Statistically significant differences at a significance level of $p < 0.05$ are illustrated by a blue background.

The statistical results confirm the high-level observations from the violin plots. For all pairwise comparisons, the hidden reference is rated significantly higher than the extrapolation methods and in all but one case (*E1*, *speech*, Fig. 5a), the extrapolation *trans* is rated significantly higher than all other extrapolation meth-



(a) *E1*, *speech* **(b)** *E2*, *speech* **(c)** *E3*, *speech*

(d) *E1*, *drums* **(e)** *E2*, *drums* **(f)** *E3*, *drums*

Fig. 5: Bonferroni-Holm-corrected p-values of all pairwise comparison tests for the extrapolation pairs *E1–E3* and *speech* or *drum* stimuli. Significant differences at $p < 0.05$ are illustrated by a blue background.

ods. The method *stat* is found significantly different from *rot* and *res* in 7 out of 12 cases. The methods *rot* and *res* are significantly different only in a single case.

4.6 Discussion

Although the assumption of the residual being constant throughout different positions in the room is not strictly met, e.g. due to the influence of room modes, the analysis of the violin plots and the statistical results clearly indicate that the extrapolation *trans* that comprises salient reflections of the target SRIR and a residual from a different position, is perceived as being more similar to the reference than the other extrapolation methods. The other methods either included no salient reflections (*res*) or salient reflection patterns that differ from the reference (*stat* and *rot*). The rendering that involved a static BRIR with dynamically rendered direct sound from the reference (*stat*) was found significantly less similar to the reference than the other methods in a majority of cases. The methods that included the rendering of a full SRIR (*rot*) or a residual SRIR (*res*) from a different position than the reference but were matched in terms of direct sound and DRR, were not found to differ significantly in their similarity to the reference. We thus conclude that the rendering of an inaccurate set of salient reflections does not contribute to perceived similarity when compared to rendering no salient reflections at all.

The most commonly reported properties that were used to rate the conditions were timbre, source direction and distance. The timbre was a few times reported to slightly differ from the reference in terms of low frequencies, which may be attributed to the influence of room modes. As the direction of the direct sound was matched with the reference in all conditions, reported shifts in direction are assumed to be caused by salient early reflections. Interestingly, many participants reported differences in perceived distance although the DRR was matched with the reference DRR in all conditions, which might be due to different ITDGs or spectral properties of the reverberation.

5 Conclusion

We conducted a listening experiment to investigate how different SRIR extrapolation methods based on a single measurement perform perceptually. Two of the methods under test involved the separation of the measured SRIR into direct part and residual using a recently proposed subspace decomposition method. The extrapolation method that performed best comprises a residual SRIR measured at a position different from the target position and salient reflections from the target SRIR. Although the method performed significantly better than the other compared methods, it was found to significantly differ from the target SRIR. However, authentic rendering, i.e., rendering that is indistinguishable from a measured reference, is often not necessary in extended reality applications. Thus, the development of a 6DoF renderer that is based on a single, measured residual in combination with modified or synthetically added salient reflections might be beneficial.

References

[1] Brinkmann, F., Aspöck, L., Ackermann, D., Lepa, S., Vorländer, M., and Weinzierl, S., "A round robin on room acoustical simulation and auralization," *The Journal of the Acoustical Society of America*, 145(4), pp. 2746–2760, 2019, doi:10.1121/1.5096178.

[2] Neidhardt, A. and Kamandi, S., "Plausibility of an approaching motion towards a virtual sound source II: In a reverberant seminar room," in *152nd Conv. Audio Eng. Soc.*, 2022.

[3] Werner, S., Klein, F., Neidhardt, A., Sloma, U., Schneiderwind, C., and Brandenburg, K., "Creation of auditory augmented reality using a position-dynamic binaural synthesis system—technical components, psychoacoustic needs, and perceptual evaluation," *Applied Sciences*, 11(3), pp. 1–20, 2021, doi:10.3390/app11031150.

[4] Wirler, S. A., Meyer-Kahlen, N., and Schlecht, S. J., "Towards transfer-plausibility for evaluating mixed reality audio in complex scenes," in *AES International Conference on Audio for Virtual and Augmented Reality (AVAR)*, 2020.

[5] Meyer-Kahlen, N., Amengual Gari, S., McKenzie, T., Schlecht, S. J., and Lokki, T., "Transfer-Plausibility of Binaural Rendering with Different Real-World References," in *Proc. of the German Annual Conference on Acoustics (DAGA)*, 2022.

[6] Arend, J. M., Amengual Garí, S. V., Schissler, C., Klein, F., and Robinson, P. W., "Six-Degrees-of-Freedom Parametric Spatial Audio Based on One Monaural Room Impulse Response," *J. Audio Eng. Soc.*, 69(7/8), pp. 557–575, 2021.

[7] Masterson, C., Kearney, G., and Boland, F., "Acoustic impulse response interpolation for multichannel systems using dynamic time warping," in *AES 35th International Conference*, 2009.

[8] Müller, K. and Zotter, F., "Auralization based on multi-perspective ambisonic room impulse responses," *Acta Acustica*, 6(25), 2020.

[9] Zhao, J., Zheng, X., Ritz, C., and Jang, D., "Interpolating the Directional Room Impulse Response for Dynamic Spatial Audio Reproduction," *Applied Sciences*, 12, 2022, doi:10.3390/app12042061.

[10] Deppisch, T., Amengual Garí, S. V., Calamia, P., and Ahrens, J., "Direct and Residual Subspace Decomposition of Spatial Room Impulse Responses," *arXiv preprint*, 2022, doi:10.48550/arxiv.2207.09733.

[11] Deppisch, T., Ahrens, J., Amengual Garí, S. V., and Calamia, P., "Spatial Subtraction of Reflections from Room Impulse Responses Measured with a Spherical Microphone Array," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 346–350, 2021, doi:10.1109/WASPAA52581.2021.9632764.

[12] McKenzie, T., McCormack, L., and Hold, C., "Dataset of Spatial Room Impulse Responses in a Variable Acoustics Room for Six Degrees-of-Freedom Rendering and Analysis (1.1)," 2022, doi:10.5281/zenodo.6382405.

[13] Schörkhuber, C., Zaunschirm, M., and Höldrich, R., "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Proc. of the German Annual Conference on Acoustics (DAGA)*, pp. 339–342, 2018.

[14] International Telecommunication Union Radiocommunication Sector, "Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," 2015.

[15] Robotham, T., Rummukainen, O. S., Kurz, M., Eckert, M., and Habets, E. A. P., "Comparing Direct and Indirect Methods of Audio Quality Evaluation in Virtual Reality Scenes of Varying Complexity," *IEEE Transactions on Visualization and Computer Graphics*, 28(5), pp. 2091–2101, 2022.

[16] Mendonça, C. and Delikaris-Manias, S., "Statistical tests with MUSHRA data," in *144th Conv. Audio Eng. Soc.*, 2018.