Audio Engineering Society

Conference Paper 8

Presented at the AES International Conference on
Headphone Technology
2025 August 27–29, Espoo, Finland

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Evaluation of room simulation approximations for AR Audio considering a flipped loudspeaker scenario

Annika Neidhardt, Boyd Thwaite, Joshua Mannall, and Enzo De Sena

University of Surrey, Guildford, UK

Correspondence should be addressed to Annika Neidhardt (a.neidhardt@surrey.ac.uk)

ABSTRACT

Spatial Audio for Augmented and Extended Reality (AR/XR) is still limited in its suitability for everyday use due to the perceptual and computational demands of such applications. A number of recent studies have highlighted the potential of simplified room acoustic modelling and its perceptual optimisation for making AR/XR applications more accessible with affordable, mobile hardware. This paper presents a perceptual evaluation of real-time acoustic modelling based on feedback delay networks (FDN), scattering delay networks (SDN), and a hybrid of the image source method (ISM) combined with FDN. The acoustics were modelled using a shoebox-approximation of the original room's geometry, two different approximations of the original loudspeaker directivity and generic HRTFs. We chose a flipped loudspeaker scenario as a critical test case. In the listening experiment, we assessed the Audiovisual Plausibility, Externalisation and Naturalness of the modelled sound fields. The SDN implementation with the subcardioid directivity was perceived similarly external and audiovisually plausible as the measured reference. The hybrid ISM-FDN method and the SDN, both with the directivity of a different loudspeaker, were perceived as natural as the reference. The tested FDN cases exhibit significantly lower ratings than the measurement for the three attributes. The chosen simplifications were not perceptually sufficient in the tested scenario, matching some attributes only. More research on efficient perceptual matching of modelled room acoustics for AR/XR is needed.

1 Introduction

A key challenge in the development of audio systems for Augmented and Extended Reality (AR/XR) is the perceptually convincing blending between virtual and real acoustic events [1]. If the target environment is known, binaural measurements can be conducted, and the data can be used for auralisation. With individualised binaural room impulse response (BRIR) measurements, it is possible to achieve a virtual simulation that cannot be distinguished from the real version [2]. Neidhardt and Zerlik [3] conducted a similar experiment for position-dynamic binaural audio. With the tested system, experienced listeners could still distinguish the virtual from the real sound source, but only if the real version was among the test items.

In AR/XR applications, achieving authenticity (i.e., the degree to which the simulation faithfully replicates the perceptual experience of a real room) is often less

important. Instead, the main objective is the creation of plausible auditory illusions with minimal computational effort.

The properties of the real sound field influence the listener's expectations towards the quality and the acoustic properties of the virtual acoustic objects. In this context, the term *transfer plausibility* was coined to describe the convincing integration of virtual acoustic objects into real acoustic environments [4]. In a related development, Gospodarek [5] evaluated plausibility of position-dynamic auralisations of measured BRIRs in comparison to real loudspeakers at different positions in the room instead of the real counterpart of the virtual content. This approach of addressing the transfer between internal and external references in a perceptual evaluation yields high ecological validity.

Recent research has started exploring the perceptual tolerances with respect to deviations of modelled sound fields from the actual sound field, for example with regard to the accuracy of early reflections [6] or the reverberation time [7]. The present study is driven by the objective to gain a better understanding of what we can "get away with" in terms of room acoustic modelling for AR/XR.

We render audio using different simplified room acoustic models. In addition to the simplifications inherent to the models we choose, the room geometry and the source directivity were also simplified, emulating the lack of precise information available in typical real-world applications. More specifically, the geometry of the real room was approximated as an empty shoebox model with similar volume. Two approximations of the real loudspeaker were tested–a measured directivity pattern of a smaller loudspeaker, and an analytical subcardioid pattern. A listening experiment evaluates the externalisation, plausibility and naturalness of the resulting position-dynamic binaural AR audio.

2 Background

This section presents a brief overview of real-time room acoustic models and their perceptual evaluation in the context of VR/XR/AR and recent efforts to create perceptually plausible simulations for AR/XR and digital-twin environments.

2.1 Room acoustic models for real-time applications

Efficient room acoustic modelling has been a subject of research for several decades. Feedback delay networks (FDN), for example, are recursive systems with multichannel delay lines where the energy is distributed using a feedback matrix [8]. Early implementations date back to the early eighties [9]. Since then, FDNs have become a popular method to simulate late reverberation.

In order to render early reflections, FDNs are commonly used in conjunction with the so-called image source method (ISM) [10]. The ISM functions by mirroring the sound source position across each enclosing surface so that every reflection can be treated as sound radiating from a virtual image source, allowing arrival times and amplitudes to be computed easily. The computational complexity of ISM increases exponentially with the reflection order, hence why they are commonly used to model only low-order reflections, which are then fed into an FDN to model the late reverberation [11, 12].

De Sena et al. [13] proposed scattering delay networks (SDN), which avoid the need for separate early reflection and late reverberation modules. SDNs consist of a digital waveguide network [14] that uses information about room geometry and surface materials to model reverberation. The system is implemented as a recursive network of delay lines connected at so-called scattering nodes, placed at the locations of first-order reflections on each wall (six in the case of a cuboid room). This results in accurate rendering of first-order reflections, while higher-order reflections are subject to progressively coarser approximations [15].

2.2 Rendering for VR/AR/XR and perceptual evaluations

A key goal of most simulation tools is to create realistic and perceptually convincing room acoustics for auralisation.

Mi et al. [16] evaluated FDN, directional FDN, Gardner, Moorer, Schroeder reverberation algorithms, a new hybrid Moorer–Schroeder reverberation algorithm in comparison with measured reverberation in a MUSHRA experiment. The evaluation was conducted using a static auralisation without head tracking. For long reverberation times, the Gardner model was rated most similar to

the measurement, while for shorter and medium reverberation times, the Schroeder and HMS reverberation algorithms were most similar.

SDNs have been shown to achieve high perceived naturalness [17], and immersion [18], and, in some cases, have been shown to result in a higher externalization than (higher-order Ambisonics-based) measurements [19].

Postma and Katz [20] showed that it is possible to generate BRIRs that are perceptually equivalent to measured BRIRs using a methodological calibration of the commercial hybrid room simulation (CATT-Acoustic) based on reverberation and clarity parameters for large venues. Brinkmann et al. [21] conducted a round robin comparison of different room acoustic simulation tools, mostly based on ray-tracing. It could be shown that the simulations still exhibit perceptual deviations from the measured impulse responses, mostly due to inaccuracies in the early reflection rendering. Kjær et al. [22] conducted a perceptual evaluation of ODEON, RAVEN and Treble, which are three room simulation models based on GA or on a hybrid of GA and numerical methods. The test used a MUSHRA elicitation methodology and static binaural auralisation. All three models produced perceptual differences from the measured reference in multiple of the attributes: Treble Strength, Bass Strength, Clarity, Level of Reverberation, Localisability, and Width.

Digital twins of real environments form a bridge between VR and AR. Stärz et al. [23], for example, created a digital twin of a real classroom to be explored in VR. In a quality assessment using a MUSHRA-like paradigm, they compared head-tracked binaural auralizations based on simulated and measured BRIRs to a loudspeaker reproduction in the real room. Simulated and measured BRIRs achieved ratings comparable with that of the real room. This holds even for simulations with generic HRTFs. They found that simulated BRIRs were rated higher than the real loudspeaker, which is a similar effect as reported by Djordjević et al. [17]. Roßkopf et al. [24] used the same digital twin classroom and examined the role of measured BRIRs, and BRIRs simulated with generic and individual HRTFs in sound localisation and social presence. While the sound localisation was comparable for all the methods, for social presence, simulations based on generic HRTFs were inferior.

Martin et al. [25] studied the perceptual implications of simplified geometry and materials characterisation. They showed that using a lower frequency resolution of absorption coefficients can influence the perception of reverberation significantly, while decimating the geometry of the model played a minor role in perception. In the context of AR, Bona et al. [26] proposed a method for automatically optimizing an artificial reverberator based on a Bayesian optimisation and implemented it for FDN and a Schroeder-Moorer reverberator. A MUSHRA-listening test suggested that the resulting reverberation does not achieve authenticity. Fantini et al. [27] examined the Co-Immersion achieved with automatically matched reverberation modelled with FDN, SDN, and other models in an Augmented Virtuality environment, a virtual environment consisting of auditory content captured from the real world, augmented by synthetic sound generation. Detection rates in a virtual vs. real paradigm were close to chance level.

In summary, several aspects of perception of (modelled) room acoustics have been studied in the context of VR. On the other hand, the perceptual effects and requirements in the context of AR are not yet fully understood. This paper contributes to filling this gap.

3 Methodology

To assess the perceptual suitability of simplified room acoustic modelling, we measured a reference BRIR data set in an ordinary seminar room. We then modelled the same scenario with three different room acoustic modelling approaches. In a psychoacoustic Absolute Category Rating (ACR) with a single stimulus test design, the modelled scenes were evaluated along the hidden measured reference.

We chose a flipped loudspeaker scenario, where the loudspeaker are facing away from the potential listening positions. This leads to cases with low direct-to-reverberant-energy ratios, which have been shown to be more critical with regard to the perception of room acoustic properties and accuracy [6].

This section explains the different steps in detail.



Fig. 1: Photo of the setup in the LTC seminar room at the University of Surrey with the G.R.A.S KEMAR 45bb HATS placed on an Outline Electronic turntable and two loudspeakers Neumann KH 120 II as sound sources.

3.1 Measured BRIR-reference data set

For the measurement of a reference BRIR data set, a seminar room called 'LTC' at the University of Surrey was chosen. The room is slightly L-shaped and furnished as shown in Fig. 1. Fig. 2 illustrates the geometry of the measurement setup with all source and receiver positions and provides the exact measurements. A G.R.A.S. KEMAR 45bb head- and torso simulator (HATS) was placed at nine positions at equal distances of 25 cm along a line with length of 2 m. One loudspeaker Neumann KH 120 II was positioned on an extension of that line at a distance of 1.25 m from the closest listening position, the second loudspeaker at 1.25 m distance to the right of the third measurement position. Both loudspeakers were flipped by 180°, so that they were facing away from the listening line. The ears of the KEMAR were located 162 cm above the floor to approximate the ear height of a standing person within the limits of the setup provided. The acoustic centre of the loudspeaker was placed at the same height. At each position along the line, BRIRs for each loudspeaker were measured at 2.5° increments for a full 360°. The HATS was rotated using an Outline electronic turntable.

According to the findings in [28], a positional resolution of 25 cm is not optimal. However, even a resolution of 5 cm showed noticeable inconsistencies in the sound field for slow walking and carefully listening for inconsistencies. However, in the context of experiments considering plausibility, 25 cm resolution was found to be sufficient with speech and music [6]. The same resolution was chosen here, to keep the measurement effort reasonable.

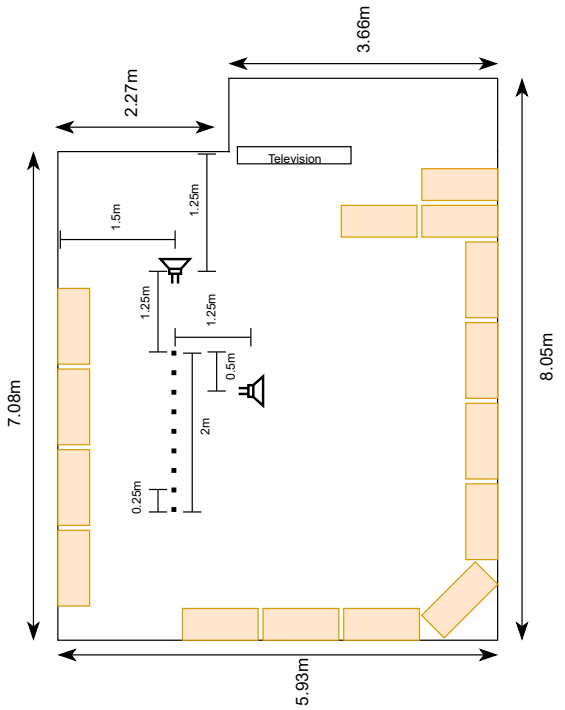


Fig. 2: Illustration of the setup in the LTC seminar room, showing all source and receiver positions used for the measurement.

3.2 Acoustic Room Simulation

The acoustic simulations are based on a shoebox approximation of the room's original shape. The simulations were conducted using the empty shoebox shown in Fig. 3 with the measures of 7 m × 6 m × 2.5 m with resulting volume V=105 m^3 and a reverberation time RT = 0.4 s (broadband). Other objects, like the furniture, were not considered. For the binaural simulation, a directional transfer function (DTF) version of the high-resolution KEMAR HRTF dataset provided by the RWTH Aachen [29] was used.

The simulations of BRIRs were implemented with the following three room acoustic modelling methods:

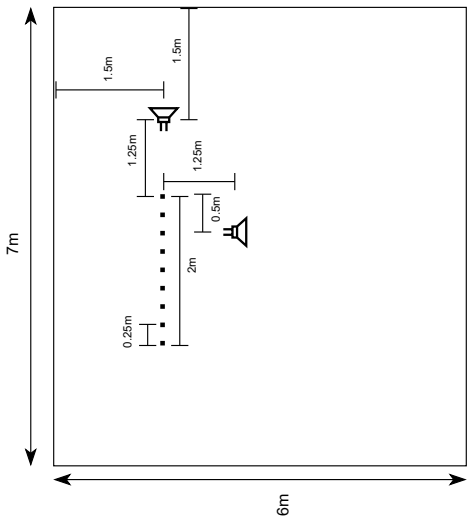


Fig. 3: Shoebox approximation of the room's shape.

FDN The FDN test cases for this study were created using the RoomAcoustiC++ toolbox [12]. A 12-channel FDN was used following the description provided by Wendt et al. [11] where the delay lines correspond to the axial dimensions of the shoebox room. The T60 was controlled by frequency-dependent filters [30] in the feedback loop of the FDN. The output was spatialised by placing reverb sources around the listener and applying HRTF processing using the 3D Tune-In toolbox (integrated into RoomAcoustiC++).

SDN The BRIRs for the SDN test cases were generated using C++ code implementing the original model in [13]. Additionally, we incorporated source directivity as described in the next subsection. The T60 in frequency bands was controlled using the same graphic equalizer used in RoomAcoustiC++ [30], applied to the delay lines at the output of each of the scattering nodes [13].

ISM-FDN This hybrid approach uses the image source method [10] for the early part of the sound field and FDN to model the late reverberation. The FDN design was the same as the above version, the difference being that the input to the FDN was the third order image sources, as in Wendt et al. [11]. The ISM-FDN test cases for this study were implemented using the RoomAcoustiC++ toolbox [12].

3.3 Directivity

Each of the room simulation methods was implemented with two different source directivity patterns to approximate the directivity of the Neumann 120 KH II loudspeaker used for the measurement. The directivity patterns used for the simulations were: a frequency-independent subcardioid directivity; and a frequency-dependent directivity made using a dataset of measurements for a Genelec 8020c, provided in the BRAS database [31].

The directivity pattern was processed as follows. First, the directivity pattern was approximated using the spherical harmonic (SH) decomposition in each frequency band. The SH order was chosen as the lowest one that resulted in an average reconstruction error of less than 0.5 dB. For 125 Hz and 250 Hz, a first order approximation was sufficient. For 500 Hz and 1000 Hz it was necessary to increase it to 2nd order. For 2000 Hz, 4000 Hz and 8000 Hz, a third order approximation was used. Second, for every rendered direction, the directivity pattern was sampled from the SH approximation and then rendered using the graphic equalizer proposed in [30]. This process was used for all three models above.

3.4 Room Acoustic Matching

Tab. 1 lists the average reverberation times estimated for the different octave bands of all the $2\times9\times144$ BRIRs measured in the room. Based on these values, the mean absorption $\overline{\alpha}$ could be determined using the Sabine equation (Eq. 1).

$$T = 0.161 \frac{V}{A} = 0.161 \frac{V}{\overline{\alpha} S} \tag{1}$$

Sabine uses an approximation of $ln(1-\overline{\alpha}) \approx -\overline{\alpha}$ which assumes that $\overline{\alpha} \ll 1$. Tab. 1 provides an overview of the mean absorption coefficients estimated based on Eq. 1. In the case of SDN, the absorption coefficients were also fine-tuned by hand to match the T60 to within ± 10 ms in each frequency band.

Table 1: Frequency-dependent T_{30}, estimated from the measured BRIRs and resulting absorption coefficients according to Sabine.

f	125	250	500	1k	2k	4k	8k
$T_{30}[s]$	0.51	0.44	0.28	0.30	0.40	0.42	0.41
α_{Sabine}	0.222	0.258	0.405	0.378	0.284	0.270	0.277

4 Listening experiment

A psychoacoustic experiment was conducted to assess the suitability of the approximation used for room acoustic modelling. The perceptually critical test scenario was designed to challenge the modelled room acoustics. The measured set of BRIRs served as a hidden reference in the experiment. Using a measured reference instead of a real one helps excluding many parameters like an imperfect headphone transparency or imperfect spatial resolution of the BRIR updates as potential bias from the investigation.

However, as the experimenter and the participant had conversations in the room prior to the experiment and throughout, we can assume that the acoustics of the real room will inform the ratings.

This section documents the setup and the test procedure in depth.

4.1 Reproduction/Test setup

The experiment was conducted in the same seminar room at the positions where the BRIRs were measured before. The body height of each participant might deviate slightly from the position in the measurement. Previous experiments had shown only minor effects and only for a few people. Both loudspeakers were still in place and the furniture was arranged as illustrated in Fig. 2.

Participants were invited to walk along the measured line of listening positions, wearing MushRoom-AR-headphones which were self-built in accordance with [32]. The proposed headphone design was developed with the goal of achieving high acoustic transparency and minimizing the effect of the headphones on the perception of the real acoustic environment. Headphone impulse responses were measured, repositioning headphones on the KEMAR head-and-torso-simulator five times and taking the average. A regularised inversion of the filter was conducted based on the least-squares approach presented in [33].

An HTC Vive Tracker 3.0 was mounted on top of the headphones to capture the motion of the listener's head with the Vive Lighthouse tracking module. All BRIRs were pre-rendered with the different tools in an equivalent configuration as the measured BRIR data set. Position-dynamic binaural rendering was implemented with the open-source tool *pyBinSim* [34], which handles and selects the BRIR filters according to the tracking-data and applies them via partitioned convolution to the dry audio signals to realise dynamic binaural synthesis in real-time.

4.2 Test items/ Test cases

Different test cases were considered for both loudspeaker positions *Front* and *Side*:

- **Meas** - Original BRIR Measurement
- **FDN**$_{sub}$ - with frequency-independent subcardioid directivity
- **FDN**$_{ls}$ - with frequency-dependent approximation of Genelec 8020c loudspeaker directivity [31]
- **SDN**$_{sub}$ - with frequency-independent subcardioid directivity
- **SDN**$_{ls}$ - with frequency-dependent approximation of Genelec 8020c loudspeaker directivity
- **ISM-FDN**$_{sub}$ - transition after 3rd order image sources - with frequency-independent subcardioid approximation of the directivity
- **ISM-FDN**$_{ls}$ - transition after 3rd order image sources - with frequency-dependent approximation of Genelec 8020c loudspeaker directivity

- **Anchor** - Originally measured BRIRs, made mono after the direct sound by copying the left channel to the right channel

Two different stimuli were used in the experiment: a solo saxophone piece and a dry male speech (excerpt of an audiobook). Each BRIR set was presented with each of the two signals to each participant in an individually randomised order.

Creating a comparable loudness of the different test scenes in a rigorous way was challenging. First, we attempted to adjust the sound level of the conditions according to the energy of their direct sound. The corresponding reverberant part contributes to the overall loudness and depends on the room modelling method. Therefore, potential changes in loudness could also be a rated feature. At the same time, the progress of the direct sound energy over the different positions and orientation varied with the test case. This raised the question, which position, orientation or which average exactly should be used as the reference direct sound level. Lacking a simple answer, we decided to adjust the sound level for this study by ear instead of physical measures.

4.3 Test method - experimental design

We conducted an ACR experiment in a single stimulus test design. The participants had to rate the auralisation along the following three questions:

- 1. Do you perceive the sound source to be outside of your head? (Externalisation)
- 2. Do you perceive the sound source to originate from the visible loudspeaker? (Audiovisual Plausibility)
- 3. Do you perceive the sound you hear as natural? (Naturalness & Quality)

Subjects had to give their answer to each of the questions using the 6-point scale shown in Tab. 2.

We chose this design over a virtual-versus-real test paradigm as proposed in [35] to avoid potential perceptual biases caused by yet not perfectly transparent reproduction systems. Furthermore, such a test method typically does not provide much information about the strengths and weaknesses of the different acoustic models, as long as they are detectably different from the real reference. However, during exploration and development, such information is valuable.

Table 2: Six-point rating scale of the experiment.

Score	Definition
1	Clearly NO
2	Mostly NO
3	Rather NO
4	Rather YES
5	Mostly YES
6	Clearly YES

4.4 Participants

A total of 21 people participated in the experiment, 19 males and 2 females. All participants are students or scientific staff in Audio Engineering and thus generally experienced listeners. Among the participants, 14 had little to no experience in listening to dynamic binaural audio for AR.

5 Results

The responses for the three evaluation questions are visualized in Fig. 4, Fig. 5 and Fig. 6, respectively. For all three questions, the ratings of the measured BRIRs show the highest mean and median, and the anchor conditions always clearly the lowest.

While a Shapiro-Wilk test rejects normality at the $\alpha = 0.05$ confidence level, the statistical analysis employed here uses parametric testing, which is generally regarded as reasonably robust to moderate departures from normality in balanced designs with a large enough sample size ($N = 84$) like our one [36]. More specifically, we use Bayesian Repeated Measures ANOVA [37] and Bayesian Related Samples T Test for the ommnibus and posthoc tests, respectively. All statistical tests were run on SPSS 29.

An advantage of Bayesian tests is that they allow to infer similarity between pairs of distributions via the Bayes Factor (BF). BF is defined as the ratio of the likelihood of the data given the hypothesis that they are from the same distribution, divided by the likelihood of the data given the hypothesis that they are not from the

same distribution. BF values above 10 are considered as strong evidence that the data is from the same distribution (in other words, that the two stimuli are similar), values between 3 and 10 are considered as moderate evidence and values between 1 and 3 are considered as anecdotal evidence [38].

The following subsections present the results of the omnibus tests run for each attribute (after having removed the anchor) and respective post-hoc analysis. The post-hoc analysis compares each of the models to Meas (6 comparisons), using Bonferroni multiple-comparisons correction.

5.1 Q1 - Externalisation

The omnibus test reveals statistically significant differences between stimuli ($p < 0.001; \chi^2 = 72.389$). Pairwise tests show that there is a statistically significant difference between nearly all models and Meas. More specifically, there are statistically significant differences between FDN$_{sub}$ and Meas (adjusted p<0.001), between FDN$_{ls}$ and Meas (adjust p<0.001), between ISM-FDN$_{sub}$ and Meas (adjusted p<0.001), between ISM-FDN$_{ls}$ and Meas (adjusted p=0.010) and between SDN$_{ls}$ and Meas (adjusted p<0.001). Vice versa, BF values indicate moderate evidence that SDN$_{sub}$ and Meas arefrom the same distribution (BF=5.484).

5.2 Q2 - Audiovisual Plausibility

The omnibus test reveals statistically significant differences between stimuli ($p < 0.001; \chi^2 = 67.875$). Pairwise tests show that there is a statistically significant difference between FDN$_{sub}$ and Meas (adjusted p<0.001), between FDN$_{ls}$ and Meas (adjust p<0.001), and between ISM-FDN$_{sub}$ and Meas (adjusted p<0.001). Vice versa, BF values indicate moderate evidence that SDN$_{sub}$ and Meas are from the same distribution (BF=5.460). The difference between ISM-FDN$_{ls}$ and Meas is borderline significant (adjusted p=0.071; unadjusted p=0.012) and so is the difference between SDN$_{ls}$ and Meas (adjusted p=0.057; unadjusted p=0.001).

5.3 Q3 - Naturalness

The omnibus test reveals statistically significant differences between stimuli ($p < 0.001; \chi^2 = 73.412$). Pairwise tests show that there is a statistically significant difference between FDN$_{sub}$ and Meas (adjusted p<0.001), between FDN$_{ls}$ and Meas (adjusted p=0.013), and between SDN$_{sub}$ and Meas (adjust p=0.019). Vice versa, BF values indicate moderate evidence that ISM-FDN$_{ls}$ and Meas are from the same distribution (BF=3.485) and anecdotal evidence that SDN$_{ls}$ and Meas are from the same distribution (BF=2.054). The difference between ISM-FDN$_{sub}$ and Meas is borderline significant (adjusted p=0.082; unadjusted p=0.014).

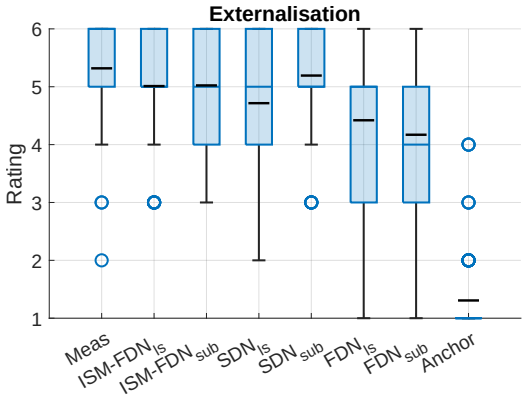


Fig. 4: Results for Question 1 - Externalisation, pooled over both source positions and both test stimuli. In addition to the medians in the boxplots, the mean values are shown in black.

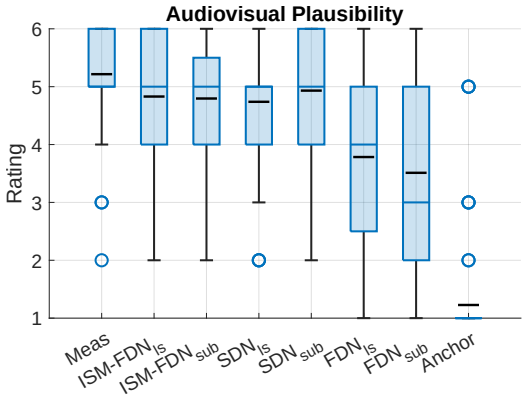


Fig. 5: Results for Question 2 - Audiovisual Plausibility, pooled over both source positions and both test stimuli. In addition to the medians indicated in the boxplot, the mean values are shown in black.

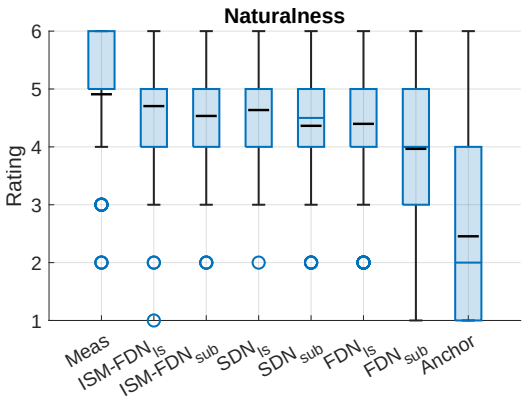


Fig. 6: Results for Question 3 - Naturalness, pooled over both source positions and both test stimuli. In addition to the medians indicated in the boxplot, the mean values are shown in black.

6 Discussion

No single room-acoustic model showed evidence of similarity to the BRIR measurements across all three perceptual attributes simultaneously. However, ISM-FDN$_{ls}$ was similar to the measurements in terms of Naturalness, SDN$_{sub}$ in terms of Externalisation and Audiovisual Plausibility, and SDN$_{ls}$ in terms of Naturalness.

Between the two SDN rendering methods, the one using the analytic subcardioid directivity (SDN$_{sub}$ seemed preferred over the one using the measured loudspeaker directivity (SDN$_{ls}$) in terms of Externalisation and Audiovisual Plausibility, which is surprising and requires further investigation.

Both FDN cases exhibited significantly lower ratings for all the attributes in the pooled analysis, and lower than the two ISM-FDN cases, confirming the importance of modelling early reflections in AR rendering. Informal comments by the subjects suggest that the source was localised closer than the loudspeaker, which is probably caused by a deviating progress of the DRR. Similar effects were reported by Gospodarek [5]. This does not necessarily imply that FDN is not capable of achieving good perceptual results in AR scenarios on its own. For example, with an automatic adjustment of the parameters to the measured data as proposed by Mezza et al. [39], FDN is likely to perform better.

6.1 Limitations and future work

In an informal preliminary listening session, three experienced listeners assessed the test cases perceptually and agreed that the measured references sounded very natural and convincing. Furthermore, although some modelling approaches created good illusions, they still sounded quite different from the measured BRIRs. However, different does not mean less plausible. As already reported by Djordjević et al. [17] and Stärz et al. [40], room acoustic simulations are sometimes even rated better than measured acoustics.

The three experienced listeners also found that out of all models, ISM-FDN$_{ls}$ seemed closest to the measurement. This was not observed in the formal experiment, which may be due to the test design intentionally concealing which stimulus was the measured reference. Further research is needed to test this hypothesis.

We conducted an ACR test with a single stimulus test design. This limited the test's sensitivity to minor differences. A paired comparison would allow for a more critical assessment. Additionally, we experimented with an atypical scenario. Most participants were not experienced with binaural audio for AR and likely did not have an accurate internal reference for the flipped loudspeaker case. This contributed to the relatively gentle ratings for modelling conditions compared to the reference. A more sophisticated training procedure before the test could reduce the effect.

To evaluate the applied simplification, a limitation of the study is that more than one simplification was applied in each test case. This makes it difficult to make clear statements about that specific type of simplification. This investigation is an exploratory study. Besides the approximations inherent in the different modelling approaches, we chose the shoebox approximation and the simplified source directivity because they yield a high potential for reducing computational load.

Another relevant future question is the development of a suitable standard method, or at least agreed-upon guidelines, for adjusting the loudness of position-dynamic scenes in psychoacoustic studies.

7 Summary and Conclusions

This paper presents a psychoacoustic experiment evaluating the suitability of selected simplifications in room acoustic modelling for Augmented and Extended Reality. A challenging AR test scenario was created with two loudspeakers positioned to face away from the listener. The original room geometry was approximated with a simple shoebox model of similar size. The influence of furnishing and two persons in the room was disregarded in the modelling process. Furthermore, the directivity of the loudspeaker used for the measurement was approximated by an analytic subcardioid pattern and, in addition, by a measured directivity for a different, slightly smaller loudspeaker. All room acoustic modelling used generic KEMAR-HRTFs.

In the listening experiment, 21 participants evaluated the Externalisation, Audiovisual Plausibility and Naturalness along three corresponding questions designed for the experiment. The results show that the modelled test cases partly achieve ratings similar to the measurement. The SDN_{sub} implementation with the subcardioid directivity was rated similar in Externalisation and Audiovisual Plausibility. ISM-FDN_{ls} and SDN_{ls} were rated as comparably natural as the measured reference. Both FDN test cases were rated significantly different for the three attributes.

We conclude that, although the frequency-dependent reverberation time was closely matched to the measured data, the simulated scenarios with the selected approximations were not always perceptually convincing, matching some attributes only. The approximations were chosen due to their practical relevance. However, the psychoacoustic responses suggest that they were not yet perceptually sufficient.

The results of this study should not be interpreted in terms of the potential of the different modelling methods. Indeed, with careful tuning of the parameters, ISM-FDN, SDN, and also FDN may perform much better. These remain open tasks for future investigations.

Acknowledgement

This work was funded by the Engineering and Physical Sciences Research Council (EPSRC) under the 'SCalable Room Acoustics Modelling (SCReAM)' grant EP/V002554/1 and the 'Challenges in Immersive Audio Technology (CIAT)' grant EP/X032914/1.

References

[1] Neidhardt, A., Schneiderwind, C., and Klein, F., "Perceptual Matching of Room Acoustics for Auditory Augmented Reality in Small Rooms - Literature Review and Theoretical Framework," *Trends in Hearing*, 26, 2022, doi:10.1177/23312165221092919.

[2] Brinkmann, F., Lindau, A., and Weinzierl, S., "On the Authenticity of Individual Dynamic Binaural Synthesis," *J. Acoust. Soc. Am.*, 142(4), pp. 1784–1795, 2017, ISSN 0001-4966, doi:10.1121/1.5005606.

[3] Neidhardt, A. and Zerlik, A. M., "The availability of a real hidden reference affects the plausibility of position-dynamic auditory AR," *Frontiers in VR*, 2021, doi:10.3389/frvir.2021.678875.

[4] Meyer-Kahlen, N., Garí, S., McKenzie, T., Schlecht, S., and Lokki, T., "Transfer-Plausibility of Binaural Rendering with Different Real-World References," in *Fortschritte der Akustik*, German Acoustical Society (DEGA), 2022.

[5] Gospodarek, M., *Acoustic and perceptual factors affecting plausibility in sound design for audio augmented reality experiences*, Ph.D. thesis, New York University, 2024.

[6] Neidhardt, A., *On the plausibility of simplified acoustic room representations for listener translation in dynamic binaural auralizations*, Ph.D. thesis, Technische Universität Ilmenau, 2023, doi:10.22032/dbt.57596.

[7] Schneiderwind, C., Richter, M., Merten, N., and Neidhardt, A., "Effects of Modified Late Reverberation on Audio-Visual Plausibility and Externalization in AR," in *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–9, 2023, doi:10.1109/I3DA57090.2023.10289186.

[8] Schlecht, S. J. and Habets, E. A. P., "On lossless feedback delay networks," *IEEE Trans. on Signal Processing*, 65(6), pp. 1554–1564, 2016.

[9] Stautner, J. and Puckette, M., "Designing multichannel reverberators," *Computer Music Journal, vol. 6, no. 1, pp. 52-65*, 1982.

[10] Allen, J. B. and Berkley, D. A., "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, 65(4), pp. 943–950, 1979.

[11] Wendt, T., van de Par, S., and Ewert, S., "A Computationally-Efficient and Perceptually- Plausible Algorithm for Binaural Room Impulse Response Simulation," *J. Audio Eng. Soc., pp. 748–766*, 2014.

[12] Mannall, J., Savioja, L., Neidhardt, A., Mason, R., and De Sena, E., "RoomAcoustiC++: An open-source room acoustic model for real-time audio simulations," in *AES Int. Conference on Headphones, Espoo, Finland*, 2025.

[13] De Sena, E., Hacıhabiboğlu, H., Cvetković, Z., and Smith III, J. O., "Efficient Synthesis of Room Acoustics via Scattering Delay Networks," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2015.

[14] Smith, J. O., "A new approach to digital reverberation using closed waveguide networks," in *Proc. of International Computer Music Conference*, pp. 47–53, 1985.

[15] Hacıhabiboğlu, H., De Sena, E., Cvetković, Z., Johnston, J., and Smith III, J. O., "Perceptual Spatial Audio Recording, Simulation, and Rendering: An Overview of Spatial-Audio Techniques Based on Psychoacoustics," *IEEE Signal Processing Magazine*, 34(3), pp. 36–54, 2017.

[16] Mi, H., Kearney, G., and Daffern, H., "Perceptual similarities between artificial reverberation algorithms," *Applied Sciences*, 2023, ISSN 2076-3417, doi:10.3390/app13020840.

[17] Djordjević, S., Hacihabiboğlu, H., Cvetković, Z., and De Sena, E., "Evaluation of the Perceived Naturalness of Artificial Reverberation Algorithms," in *148th AES Convention, Paper 10353, Online*, 2020.

[18] Potter, T., Cvetković, Z., and De Sena, E., "On the Relative Importance of Visual and Spatial Audio Rendering on VR Immersion," *Frontiers in Signal Processing*, 2, 2022, ISSN 2673-8198.

[19] Geronazzo, M., Tissieres, J. Y., and Serafin, S., "A minimal personalization of dynamic binaural synthesis with mixed structural modeling and scattering delay networks," in *Proceedings of the 2020 International Conference on Acoustics, Speech and Signal Processing*, 2020.

[20] Postma, B. N. J. and Katz, B. F. G., "Perceptive and objective evaluation of calibrated room acoustic simulation auralizations," *J. Acoust. Soc. Am.*, 140(6), pp. 4326–4337, 2016, ISSN 0001-4966, doi:10.1121/1.4971422.

[21] Brinkmann, F., Aspöck, L., Ackermann, D., Lepa, S., Vorländer, M., and Weinzierl, S., "A round robin on room acoustical simulation and auralization," *J. Acoust. Soc. Am.*, 145(4), pp. 2746–2760, 2019, doi:10.1121/1.5096178.

[22] Kjær, C., Volk, C. P., and Jeong, C.-H., "Perceptual Evaluation of Room Acoustic Simulations and Measurements," in *Proceedings of Baltic-Nordic Acoustic Meeting, pp. 255-262*, 2024.

[23] Stärz, F., Kroczek, L., Roßkopf, S., Mühlberger, A., van De Par, S., and Blau, M., "Comparing Room Acoustical Ratings in an Interactive Virtual Environment to Those in the Real Room," in *Forum Acusticum Torino, Italy*, 2023, doi: 10.61782/fa.2023.0525.

[24] Roßkopf, S., Kroczek, L. O., Stärz, F., Blau, M., Van de Par, S., and Mühlberger, A., "The impact of binaural auralizations on sound source localization and social presence in audiovisual virtual reality: converging evidence from placement and eye-tracking paradigms," *Acta Acust.*, 8, p. 72, 2024, doi:10.1051/aacus/2024064.

[25] Martin, V., Engel, I., and Picinali, L., "Perceptual implications of simplifying geometrical acoustics models for Ambisonics-based binaural reverberation," *arXiv, 2411.10375v1*, 2024.

[26] Bona, R., Fantini, D., Presti, G., Tiraboschi, M., Engel Alonso-Martinez, J. I., and Avanzini, F., "Automatic Parameters Tuning of Late Reverberation Algorithms for Audio Augmented Reality," in *Proc. of 17th International Audio Mostly Conference*, p. 36–43, 2022, doi:10.1145/3561212.3561236.

[27] Fantini, D., Presti, G., Geronazzo, M., Bona, R., Privitera, A. G., and Avanzini, F., "Co-immersion in Audio Augmented Virtuality: the Case Study of a Static and Approximated Late Reverberation Algorithm," *IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 11, pp. 4472-4482.*, 2023, doi:10.1109/TVCG.2023.3320213.

[28] Neidhardt, A. and Reif, B., "Minimum BRIR grid resolution for interactive position changes in dynamic binaural synthesis," in *148th AES Convention, paper 10371, Vienna, Austria*, 2020.

[29] Braren, H. and Fels, J., "A High-Resolution Head-Related Transfer Function Data Set and 3D-Scan of KEMAR," *RWTH Publications*, 2020, doi:10.18154/RWTH-2020-11307.

[30] Oliver, R. J. and Jot, J.-M., "Efficient multi-band digital audio graphic equalizer with accurate frequency response control," in *139th AES Convention, New York, NY, USA*, 2015.

[31] Aspöck, L., Brinkmann, F., Ackermann, D., Weinzierl, S., and Vorländer, M., "BRAS - Benchmark for Room Acoustical Simulation," 2020, doi:10.14279/depositonce-6726.3.

[32] Mülleder, A., Romanov, M., Meyer-Kahlen, N., and Zotter, F., "Do-it-yourself headphones and development platform for augmented-reality audio," in *AES International Conference, Huddersfield, UK*, 2023.

[33] Schärer, Z. and Lindau, A., "Evaluation of equalization methods for binaural signals," in *126th AES Convention, paper 7721, Munich, Germany*, 2009.

[34] Neidhardt, A., Klein, F., Knoop, N., and Köllmer, T., "Flexible Python Tool for Dynamic Binaural Synthesis Applications," in *142nd AES Convention, Berlin, Germany*, 2017.

[35] Lindau, A. and Weinzierl, S., "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acustica united with Acustica*, 98(5), pp. 804–810, 2012, ISSN 16101928, doi:10.3813/AAA.918562, doi: 10.3813/AAA.918562.

[36] Glass, G. V., Peckham, P. D., and Sanders, J. R., "Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance," *Review of Educational Research*, 42(3), pp. 237–288, 1972.

[37] Rouder, J. N., Morey, R. D., Speckman, P. L., and Province, J. M., "Default Bayes factors for ANOVA designs," *Journal of Mathematical Psychology*, 56(5), pp. 356–374, 2012.

[38] Kass, R. E. and Raftery, A. E., "Bayes Factors," *Journal of the American Statistical Association, Vol. 90, No. 430, 773-795*, 1995.

[39] Mezza, A., Giampiccolo, R., De Sena, E., et al., "Data-driven room acoustic modeling via differentiable feedback delay networks with learnable delay lines," *J. Audio Speech Music Proc., 51*, 2024, doi:10.1186/s13636-024-00371-5.

[40] Stärz, F., Kroczek, L., Roßkopf, S., Mühlberger, A., van De Par, S., and Blau, M., "Perceptual comparison between the real and the auralized room when being presented with congruent visual stimuli via a head-mounted display," in *Proc. 24th International Congress on Acoustics, Gyeongju (Korea)*, 2022.