PERCEPTUAL COMPARISON OF AMBISONICS-BASED REVERBERATION METHODS IN BINAURAL LISTENING

Isaac Engel[1] **Craig Henry**[1] **Sebastià V. Amengual Garí**[2]
Philip Robinson[2] **David Poirier-Quinot**[3] **Lorenzo Picinali**[1]
[1] Dyson School of Design Engineering, Imperial College London, United Kingdom
[2] Facebook Reality Labs, Redmond, US
[3] Sorbonne Université, CNRS, Institut Jean Le Rond d'Alembert, France
isaac.engel@imperial.ac.uk

ABSTRACT

Reverberation plays a fundamental role in the auralisation of enclosed spaces as it contributes to the realism and immersiveness of virtual 3D sound scenes. However, rigorous simulation of interactive room acoustics is computationally expensive, and it is common practice to use simplified models at the cost of accuracy. In the present study, two subjective listening tests were carried out to explore trade-offs between algorithmic complexity (and approach) and perceived spatialisation quality in a binaural spatialisation context. The first experiment assessed the perceived realism of room reverberation, comparing auralisations based on Ambisonic impulse responses at varying resolutions (zeroth to fourth order). The second experiment focused on the perceptual relevance of different approaches to binaural reverb rendering, looking at statically or dynamically rendered room simulations. Throughout these experiments, the direct sound path was rendered separately by convolution with a Head Related Impulse Response (HRIR). Preliminary results suggest that, for the conditions under test, there may not be perceivable benefits in using high order Ambisonics encoding (beyond first order) for room auralisation and that introducing head-tracking may have little impact as well, as long as the direct sound is rendered with enough accuracy. Further work is outlined with regards to continuing this research with a higher number of participants and more varied tested conditions to clarify the extent to which these conclusions can be made.

1. INTRODUCTION

In the years since digital reverb was first suggested by Schroeder and Logan [1] there has been continual development and improvement in the technology. Until recently, research has been driven predominantly by two industries: acoustic architecture and music, in which offline computation is generally sufficient. However, the tech industry is responding to the ever-growing demand for interactive audiovisual experiences. Dynamic room auralisation is an important aspect of mixed reality experiences as it contributes to the quality of spatialisation [2]. With the recent emergence of portable Virtual and Augmented Reality, which have a limited computational power compared to desktop computing, lightweight room auralisation and acoustic modelling are essential. It is therefore important to establish the perceptual relevance of different spatial, temporal and spectral attributes of acoustic room responses to identify where computational savings can be made.

2. BACKGROUND AND MOTIVATIONS

Reverberation is the result of the pairing of an acoustic source with an environment. As the wave propagates from the source it interacts with its surroundings. The geometry of the space causes the wave to reflect and diffract. As a result, the wave arrives at the receiver via different paths. Sound has a sufficiently slow speed that these arrivals occur at perceptually distinct time intervals. As the wave continues to interact with the room, higher order reflections decrease in amplitude and the echo density increases, eventually resulting in a diffuse reverberant sound field. A room impulse response (RIR) is typically decomposed into three sections: direct sound, early reflections and late reverberation. The geometry of the room and acoustic properties of the materials therein affect a number of objective acoustic characteristics.

Drawing parallels between objective and subjective spatial features has long been the aim of different avenues of acoustic research. Current literature states that: (i) Direct sound reaches the listener first, allowing the listener to localise the source. (ii) However, strong specular reflections arriving within 5-10ms can contribute to creating a perceived image shift and broaden the apparent source width [3]. (iii) In addition, they colour the timbre of the direct sound due to phase cancellations and subsequent comb-filtering [4]. These reflections are not localised as a separate event due to the precedence effect [5]. (iv) Kaplanis et al. [6] state that the time delay between the direct sound and the first perceptually distinct early reflection affects the perception of presence and the environment dimensions, and that (v) the temporal characteristics of subsequent early reflections also contribute to the environment size. (vi) The timing, direction and spectrum of early lateral reflections contribute to the envelopment of the space [7]. As the density of the reflections increases, perception is governed less by temporal characteristic and more by statistical properties of the reverberant tail. (vii) Research done by Yadav et al. [8] suggests that the reverberation time contributes to the perception of size most significantly in large rooms, whereas early reflections are of greater importance in small rooms. Further observations have been made with respect to binaural rendering. (viii) Reverberation improves the externalisation of sound sources in the binaural domain, even if only early reflections are used [2]. (ix) The congruence between the room acoustic properties of the listening space and those of the presented virtual sounds affect the level of externalisation [9].

When modelling these features for a real time interactive environment, it is the early reflections which present the greatest difficulty, as literature suggests that in order to realistically render a room, each early reflection must be properly spatialised. Otherwise, perceptually relevant characteristics such as envelopment, room dimensions and presence would be inaccurately represented in the resultant signal. On the contrary, the reverberant tail is characterised by statistical properties, individual reflections cannot be differentiated and are uniformly distributed around the listener in their direction of arrival.

Rigorous rendering of head-tracked early reflections generally has a high computational cost. There are two primary solutions: convolution-based reverb and algorithmic reverb. The former relies on convolution with pre-measured RIRs which are dynamically swapped with head movements. Whereas, algorithmic reverb attempts to approximate the reverb by various means [10] [11] [12] and, when rendered dynamically, filters must be recomputed in real time. In practice it is common to use simplified models at the cost of accuracy. Therefore, it is relevant to investigate the extent to which computations can be simplified (e.g. by reducing the spatial resolution) without losing perceived spatialisation quality. This paper will concentrate on convolution-based methods, however, conclusions may also be relevant in algorithmic reverb.

3. PREVIOUS WORK

Prior to this paper a study was carried out assessing the realism and localisation accuracy achieved with five different binaural reverberation rendering methods [13]. The primary method was an approach which will be referred to as Reverberant Virtual Loudspeaker decoding (RVL) method.

3.1 RVL: an efficient binaural reverberation method

The RVL method is used to efficiently produce dynamic convolution-based binaural reverberation by using a set of Binaural Room Impulse Responses (BRIRs). The method is inspired by the classic virtual loudspeaker approach, first outlined by McKeag and McGrath [14] and later used by Noisternig et al. [15]. In the original method, one or more sound sources are encoded in the Ambisonic domain, which is then decoded to a set of virtual loudspeakers distributed around the listener. The binaural signal is then obtained by convolving each loudspeaker feed with the HRIR corresponding to its location. RVL follows the same approach, however it uses BRIRs in the loudspeaker position instead of HRIRs, effectively integrating the acoustics of the room in the binaural rendering.

This method has two limitations. The first being that the early reflections for a given source-receiver pair are approximated by the reflections of the limited set of loudspeaker positions. The second is that, when using dynamic rendering, the relative position of the sound sources can be changed in the Ambisonics domain, but the room is head-locked due to the set of BRIRs being fixed. This means that a rotation of the listener's head is equivalent to all of the sources in the room rotating in the opposite direction. This would be irrelevant in an anechoic scenario (traditional virtual loudspeaker method) but not in this case.

The main advantage of the RVL method over other convolution-based approaches is that, as all sound sources are encoded in the same Ambisonic soundfield, the number of required convolutions is independent from the number of sources. This is a big advantage when several sources need to be dynamically rendered at the same time. Furthermore, because the decoding process from Ambisonic channel to the left and right binaural channels is linear and time-invariant, the direct transfer functions between them can be calculated, reducing the total number of convolutions needed in real time to 2 per Ambisonic channel (e.g. 8 convolutions for first order, 18 for second order, etc.). These simplifications create a significant computational saving over other methods (convolution-based or otherwise), as the room auralisation is inherent in the rendering process, instead of having to rely on additional steps to incorporate reverb. Whether the simplifications have a negative impact on the perceived quality of spatialisation, is the question that will be addressed in the present study.

3.2 Previous study: method and results

In the previous study, a subjective listening test was carried out to evaluate the impact of the spatial resolution of reverberation in the perceived quality of binaural rendering [13]. The tested reverberation methods ranged from higher resolution and complexity (RVL with third order Ambisonics and 20 virtual loudspeakers) to less complex solutions (stereo). It also included simplistic approaches such as diotic monophonic reverb and 'mono panned', where the reverb of each sound source was added to the direct sound and spatialised in the same position. For all tested conditions, the direct sound path was rendered by convolving the dry audio with an anechoic HRIR, and was identical across conditions. The BRIRs for the different virtual loudspeaker positions were simulated by means of geometrical acoustic modelling software, without computing the

doi:10.25836/sasp.2019.11

direct sound path.

In the listening test, subjects were presented with all possible pairs of conditions and were asked to compare them in terms of plausibility and spatialisation accuracy. The audio scene included a female voice and footstep sounds moving slowly around the listener. The test was implemented in a online platform in order to reach a large population (75 participants). Results showed that no significant differences were present between four of the five tested methods, with 'mono panned' performing the worst. This seems to suggest that the level of complexity in the reverberation method does not always yield perceptually relevant improvements.

Due to this early test being web-based, this study had some limitations, such as the renderings being static (no head tracking) and the lack of control over the headphones being used by the participants and the playback level. As such, broader conclusions were unable to be made unless these issues were resolved.

4. CURRENT EXPERIMENTS

The literature review in section 2 suggests that early reflections must be properly spatialised to recreate perceptually relevant room characteristics. However the study described in section 3 suggests that increased algorithmic complexity above a certain threshold may no longer lead to improved perceived realism.

As opposed to the aforementioned web-based approach, the current study aimed to carry out experiments under laboratory conditions with dynamic (head-tracked) binaural rendering. Furthermore, reverberation based on recorded IRs is employed instead of geometrical modelling.

Two experiments are outlined in this paper, both of which were designed to test the perceptual limits of binaural rendering complexity. The methods used reproduce early reflections and tails with differing levels of accuracy to test participants' ability to discern between them and state subjective preference on the basis of perceived realism. The questions that these experiments try to answer are:

1. What is the perceptual impact of decreasing the Ambisonic order of reverberation, if direct sound is still rendered accurately? (first experiment)

2. What is the perceptual impact of rendering reverberation statically or dynamically, and using a simplified approach such as RVL? (second experiment)

4.1 Measurements

A small meeting room (RT = 900 ms) was measured for this study in two different ways: (i) RIRs were recorded with an Eigenmike microphone (MH Acoustics [1]) from three directions (-30/0/30° azimuth, 0° elevation) and then encoded into first-fourth order Ambisonics using the EigenStudio software package. (ii) BRIRs were recorded with a KEMAR head and torso simulator from six directions (front, back, left, right, up, down). All measurements were made using the sine sweep technique [16] and a Genelec 8030 loudspeaker at a distance of 1.2 m.

Figure 1: Picture of the room used in this study, before measuring RIRs with the Eigenmike.

The key to this study was that the direct sound path was always rendered separately from the reverberation, which allowed for individual control. To that end, all RIRs and BRIRs had the direct sound path removed by replacing the first 3.71 ms after the onset of all impulse responses with silence and applying a Hanning window – this time was calculated by subtracting the direct sound propagation time from the first reflection (floor) propagation time, minus a safety window of 30 samples (0.68 ms).

Given that the direct sound was obtained from a separate system and database (see next subsection), all impulse responses had to be normalised to match the direct-to-reverberant ratio of the actual room. Furthermore, the Ambisonic RIRs were equalised with a second order low-shelf filter at a gain of -15 dB and a cutoff frequency of 1 kHz to match the perceived coloration of a KEMAR BRIR.

4.2 Audio material and binaural rendering

The stimulus used in the test was an extract of female English speech from the 'Music for Archimedes' corpus [2]. The sound source was spatialised at -30° azimuth, 0° elevation and a distance of 1.2 m.

The direct sound was always rendered through convolution with the Head Related Impulse Response (HRIR) of a KEMAR from the SADIE II database [3]. To generate the reverberation, two different methods were used:

1. **Ambisonic reverb**, convolving the source with the Ambisonic RIRs, and decoding the soundfield into virtual loudspeakers which are then rendered binaurally using HRTFs.

2. **Reverberant Virtual Loudspeaker decoding (RVL) reverb**, using the recorded KEMAR BRIRs.

[1] mhacoustics.com/products

[2] pcfarina.eng.unipr.it/Public/Aurora_CD/Anecoic/Archimedes/CD-cover/Archimedes.htm

[3] york.ac.uk/sadie-project/database.html

The 3D Tune-In Toolkit [17] was used as the spatial audio engine. It binaurally rendered both the direct sound and virtual loudspeakers, and allowed for head-tracking using an EdTracker Pro Wireless [4] .

4.3 Experiment 1: MUSHRA test

One of the ways to reduce the number of convolutions required in Ambisonics-based reverb is to decrease the Ambisonic order. However, this has been shown to deteriorate localisation of direct sources [18] [19]. It was the aim of this experiment to assess whether this degraded spatialisation accuracy was also perceptually relevant to the reverberation process, with the direct sound being rendered through a full set of HRIRs.

Listeners were asked to rate the quality of six conditions where reverberation was rendered using differing Ambisonics orders and virtual loudspeaker configurations:

- **Dry** – anechoic direct sound.
- Zeroth order Ambisonics (**0OA**) with 6 virtual loudspeakers in a tetrahedron setup, each playing the W channel.
- First order Ambisonics (**1OA**) with 6 virtual loudspeakers in a tetrahedron setup.
- Second order Ambisonics (**2OA**) with 12 virtual loudspeakers in an icosahedron setup.
- Third order Ambisonics (**3OA**) with 20 virtual loudspeakers in a dodecahedron setup.
- Fourth order Ambisonics (**4OA**) with 32 virtual loudspeakers in a Pentakis-dodecahedron setup.

The six stimuli were presented in a MUSHRA listening test format [20] implemented in MaxMSP, with the **4OA** method acting as the reference, and the **Dry** stimuli as the anchor. Subjects were asked to rate the similarity of each stimulus to the reference on a scale from 0 to 100, where the latter means 'identical to the reference'. In addition, the user interface showed a picture of the simulated room and a diagram with the relative position of the sound source, to assist the subjects in creating an internal reference of the scenario being rendered. Listeners were encouraged to use head movements to explore the scene.

4.4 Experiment 2: Paired comparisons

The second experiment focused on the perceptual relevance of different approaches to binaural reverb rendering. Three different rendering methods were compared:

- Ambisonic reverb, with first order Ambisonics and 6 virtual loudspeakers – i.e. the **1OA** method from the MUSHRA test.
- **1OAS**: Static (non-head-tracked) Ambisonic reverb, with first order Ambisonics and 6 virtual loudspeakers – direct sound path was still tracked.
- First order **RVL** reverb obtained from the KEMAR BRIRs, which is inherently "static" in the sense that the room is head-locked as explained in section 3.1.

The two key factors explored in the second experiment are (i) the importance of the sound field being rendered statically or dynamically and (ii) the perceptual relevance of accurate early reflections. The three tested methods approximate early reflections to different extents by making various simplifications. In theory, this should interfere with most of the perceptual attributes outlined in section 2 as it alters the spatial, temporal and spectral characteristics of the room. Whether the differences between these approaches are perceptually relevant is still to be understood.

In each trial, listeners were shown a picture of the rendered room and a diagram with the position of the sound source, and were presented two stimuli, A and B. The question asked was '*Considering the given scene, which example is more appropriate?*'. The rating scale was continuous from -2 to +2, with one decimal place, from *Definitely A* to *Definitely B* (see Fig. 3). Listeners were allowed to freely switch between the synchronised stimuli during a trial, and head movements were encouraged to explore the scene. All possible pairs of the three rendering methods were tested, plus two null pairs where A and B were identical (randomly chosen), totalling 8 trials for each subject.

5. INITIAL RESULTS

At the time of writing this paper, the study is still ongoing and only preliminary data are reported. Results for the first five subjects (ages 23-40, 1 female) are presented.

5.1 MUSHRA test

The results of the MUSHRA test are shown in Fig. 2. Descriptive analysis showed that the mean rating for **Dry** results is clearly the lowest, followed by **0OA**, while the other four methods had relatively similar mean ratings.

Due to the low amount of subjects available so far, non-parametric statistical analysis was used. Friedman test showed that the differences between methods is significant ($\chi^2(5) = 20.28, p = 0.001$). At this early stage, post-hoc test results are not reported as are not likely to be reliable due to the low participant count. However, inspecting the boxplot diagram it seems clear that the significant differences are likely between **Dry** and the rest, and between **0OA** and the rest.

5.2 Paired comparisons

Figure 3 shows the comparison ratings for every pair of stimuli. Descriptive analysis showed that mean rating was close to zero for the null pairs, that listeners tended to favour **1OA** and **1OAS** over **RVL**, and that **1OA** and **1OAS** were perceived as very similar, with a slight trend towards favouring the latter.

As done for the other test, non-parametric statistical analysis was used; the Friedman test showed that the differences between the tested pairs were not significant

[4] edtrackerpro.mybigcommerce.com/edtracker-pro-wireless/

doi:10.25836/sasp.2019.11

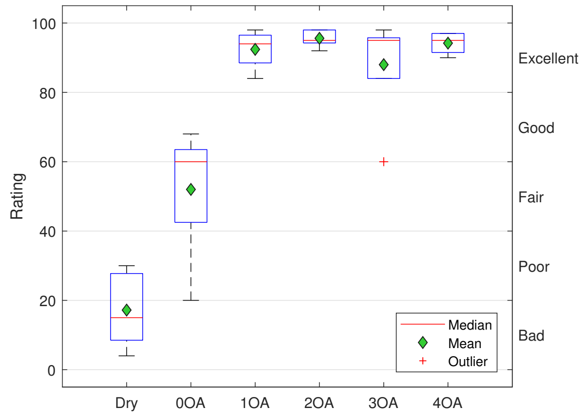

Figure 2: Boxplot diagram of the ratings for each of the tested conditions in the MUSHRA test.

($\chi^2(3) = 6.21, p = 0.102$). Furthermore, one-sample t-tests showed that none of the pairs were significantly different from a normal distribution with mean equal to zero.

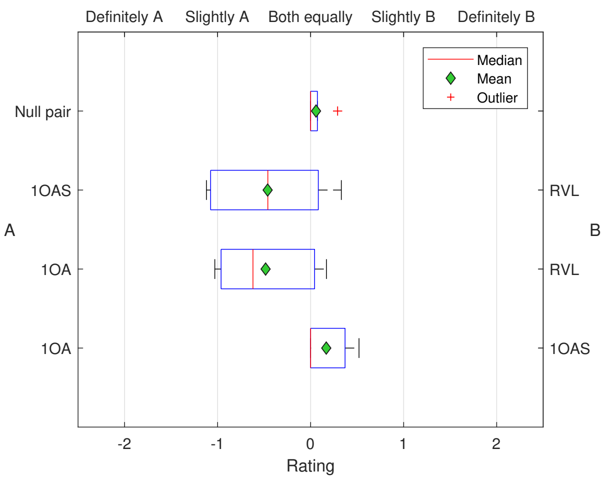

Figure 3: Boxplot diagram of the paired comparison ratings. Negative ratings indicate preference for the stimulus on the left (A), and positive ratings for the stimulus on the right (B).

6. DISCUSSION AND FURTHER WORK

This paper presents early stage results with limited scope considering the range of variables which could affect the outcome of this experiment, namely the program content (i.e. spatialised signal), room size and direct to reverberant ratio, and the participant's level of listening expertise.

With this in mind, the results from the MUSHRA test showed that while listeners gave consistently low ratings to the **Dry** and **0OA** conditions, they were not able to reliably distinguish between first order Ambisonics and above. This result suggests that the findings of [18] [19] regarding improvements in spatialisation in higher orders of Ambisonics may not extend to the reverberant portion of room auralisation, for the conditions tested so far in this study.

The implication of this would be that, provided the direct sound is rendered with sufficient accuracy, the perceptual impact of reverberant spatial resolution may not be as high as expected. Further work (currently being carried out) includes collecting data from additional participants (including a number of expert listeners) and using a range of stimulus types and room shapes and sizes.

A relevant matter that may have influenced these results is that, due to the listening test taking place outside of the measured room, room divergence is masking differences in the performance of the different Ambisonic orders [9]. However, due to the MUSHRA test focusing on perceived differences against a reference, this seems unlikely. In order to eliminate this potential variable, it will be considered to conduct future tests inside the rooms corresponding to the measured impulse.

In the second experiment, no statistically significant differences were found, but the trend suggests that listeners slightly favoured Ambisonic reverb over the RVL method. This may indicate that the simplifications made by RVL could have had a negative, yet not significant, impact in the perceived quality of the reverb. However, it could also be that timbral differences due to the methods being generated from different microphones' recordings – only partially mitigated by heuristic equalisation – were culpable instead. Additionally, listeners were exposed to the MUSHRA test where the reference was an Ambisonic reverb, and may have been biased towards this technique and its timbral quality.

Interestingly, **1OAS** was rated similarly to **1OA**. This is a surprising result, as it suggests that rendering reverb statically instead of dynamically may not be perceptually relevant, assuming that the direct path is rendered through convolution with an HRIR. This goes against considerable literature supporting the contrary, and the scope of these findings is currently limited. The number of participants is still low, and only one room and one stimulus type have been tested. More participants and more varied test conditions are needed to make such claims. Additionally, whilst listeners were encouraged to move their heads, in future work this should be considered with greater detail, either by tracking the movement of participants or prescribing specific head rotations to be carried out.

7. CONCLUSIONS

In this study, the issue of the trade-off between computational complexity and perceived quality for binaural reverberation has been addressed. Preliminary results of a perceptual listening test suggest that the Ambisonic order and the use of head tracking have little perceivable effect on Ambisonic-based reverb, assuming that the direct sound path rendering is accurate enough. Further work has been outlined in the Discussion section.

8. ACKNOWLEDGEMENTS

This work was partly supported by the PLUGGY project (https://www.pluggy-project.eu/), European Unions Horizon 2020 research and innovation programme under grant agreement No 726765.

9. REFERENCES

[1] M. R. Schroeder and B. F. Logan, "'colorless' artificial reverberation," *J. Audio Eng. Soc*, vol. 9, no. 3, pp. 192–197, 1961.

[2] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source," *Journal of the Audio Engineering Society*, vol. 49, pp. 904–916, oct 2001.

[3] S. E. Olive and F. E. Toole, "The detection of reflections in typical rooms," *Journal of the Audio Engineering Society*, vol. 37, no. 7, pp. 539–553, 1989.

[4] S. Bech, "Timbral aspects of reproduced sound in small rooms. II," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3539–3549, 1996.

[5] H. Wallach, E. B. Newman, and M. R. Rosenzweig, "The precedence effect in sound localization," *The American Journal of Psychology*, vol. 62, no. 3, pp. 315–336, 1949.

[6] N. Kaplanis, S. Bech, S. H. Jensen, and T. van Waterschoot, "Perception of reverberation in small rooms: a literature study," in *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*, pp. 1–14, Audio Engineering Society, 2014.

[7] M. Barron and A. H. Marshall, "Spatial impression due to early lateral reflections in concert halls: the derivation of a physical measure," *Journal of Sound and Vibration*, vol. 77, no. 2, pp. 211–232, 1981.

[8] M. Yadav, D. A. Cabrera, L. Miranda, W. L. Martens, D. Lee, and R. Collins, "Investigating auditory room size perception with autophonic stimuli," in *Audio Engineering Society Convention 135*, p. 10, Audio Engineering Society, 2013.

[9] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, IEEE, June 2016.

[10] J. C. Allred and A. Newhouse, "Applications of the monte carlo method to architectural acoustics," vol. 30, no. 1, pp. 1–3.

[11] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smallroom acoustics," vol. 65, no. 4, pp. 943–950.

[12] E. De Sena, H. Haciihabiboglu, Z. Cvetkovic, and J. O. Smith, "Efficient synthesis of room acoustics via scattering delay networks," vol. 23, no. 9, pp. 1478–1492.

[13] L. Picinali, A. Wallin, Y. Levtov, and D. Poirier-Quinot, "Comparative perceptual evaluation between different methods for implementing reverberation in a binaural context," in *Audio Engineering Society Convention 142*, pp. 1–13, 2017.

[14] A. Mckeag and D. Mcgrath, "Sound Field Format to Binaural Decoder with Head Tracking," *AES 6th Australian Regional Convention*, 1996.

[15] M. Noisternig, T. Musil, A. Sontacchi, and R. Holdrich, "3d binaural sound reproduction using a virtual ambisonic approach," in *IEEE International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems*, pp. 174–178, IEEE, 2003.

[16] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Audio Engineering Society Convention 122*, Audio Engineering Society, 2007.

[17] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona, "3d tune-in toolkit: An open-source library for real-time binaural spatialisation," *PLOS One*, vol. 14, no. 3, 2019.

[18] E. Bates, G. Kearney, D. Furlong, and F. Boland, "Localization accuracy of advanced spatialisation techniques in small concert halls," *The Journal of the Acoustical Society of America*, vol. 121, no. 5, pp. 3069–3070, 2013.

[19] A. Sontacchi, P. Majdak, M. Noisternig, and R. Höldrich, "Subjective Validation of Perception Properties in Binaural Sound Reproduction Systems," *AES 21st International Conferece*, pp. 1–4, 2002.

[20] "Itu-r bs. 1534-3: Method for the subjective assessment of intermediate quality level of audio systems," October 2015.

 doi:10.25836/sasp.2019.11