Proof of Concept of a Binaural Renderer with Increased Plausibility

Ulrike Sloma[1], Nils Merten[1], Thomas Thron[1], Karlheinz Brandenburg[12], Franciska Wollwert[1], Renato Profeta[1], Cristina Rodriguez[1]

[1] *Brandenburg Labs GmbH, 98693 Ilmenau, Germany, Email: us@brandenburg-labs.com*

[2] *Technische Universität Ilmenau, 98693 Ilmenau, Germany, Email: karlheinz.brandenburg@tu-ilmenau.de*

Introduction

For a long time there have been proposals for binaural rendering algorithms for headphones, trying to achieve perfect immersion. Based on basic research at TU Ilmenau, TH Köln and others, Brandenburg Labs (BLS) built a proof of concept demo showcasing the comparison of a real loudspeaker setup and headphones based rendering in a given room. It improves on previous systems by including room acoustic processing feasible to run in real time. In the past two years, this demo has been shown at a number of occasions including Tonmeistertagung, Schoeps Mikroforum and the AES Conference on Audio for Virtual and Augmented Realities. This paper introduces backgrounds of the technology and a formal listening test. It aims to verify the feedback given at the demo booths, that plausible playback via binaural rendering is possible.

Motivation

Binaural synthesis has been subject of research for way more than 40 years now and is psychoacoustically motivated. A key to the success of immersive audio technologies is the knowledge that true immersion is created inside our brain. Reproduced signals must correspond to the expectations of how acoustic events should sound. Besides several other aspects, this is especially true for the acoustics of the environment the listener is in. A mismatch can lead to degradation of immersion, the so called room acoustic divergence effect (RDE) [1]. Therefore it is necessary to incorporate those room acoustics. Another important cue is the possibility to explore the room and the present sound sources. Like in real acoustic environments, being able to traverse the scene and listen from different sides and perspectives helps to build an understanding of the sound field. This is one key to an externalized acoustic scene. To enable this a constant low latency head tracking in 6 Degrees of Freedom (6DoF) is needed. In contrast to previous research, the use of individualized Head Related Transfer Functions (HRTFs) is not mandatory for most listeners to achieve an externalized auditory impression, if sufficient cues of the acoustic scene are available. This was previously proven in research [2] and now with the immersive audio demo, which uses a generic HRTF dataset for auralization.

Immersive Audio Demo

To introduce the technology of BLS to the public, a demo setup was created (figure 1). The demo presents a fully virtual and dynamic doppelganger of a real pair of stereo speakers, auralized over headphones. It is possible to switch to their real counterparts on demand, allowing direct comparison. To prepare the demo a room acoustic measurement for each sound source has been done and room dimensions have been measured. This demo was presented at several acoustically different venues. In total, a plausible, well externalized reproduction was confirmed by more than 600 listeners now. While there are minor audible differences, e.g. in timbre, less then 1% of the listeners reported bad externalization or plausibility.

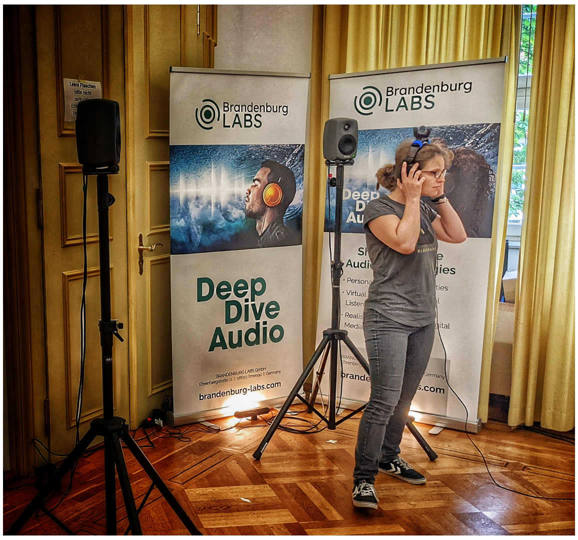

Figure 1: Immersive audio demo (Schoeps Mikroforum 2022)

Basic Algorithms and Principles

Both technologies used in the presented study are based on a broad expertise and research at TU Ilmenau [2, 3] about binaural audio reproduction and perception of auralized audio. Further results and algorithms from Aalto University and TH Köln regarding the simulation of the sound field are included in the implementations.

C3k

The algorithms used for the BLS methodology are based on algorithms described in [4] and have numerous further additions. At it's core, it is a parametric extrapolation algorithm. It calculates Binaural Room Impulse Responses (BRIRs) in real-time, based on a single omnidirectional Room Impulse Response (RIR). A very basic room geometric model as well as the positions of the sound sources and microphone need to be captured. From that, the Directions of Arrival (DoAs) of the direct sound and early reflections are estimated by a simplified image source model. The RIR is processed in segments and appropriately convolved with generic HRTF filters. Late reverberation is simulated by noise shaping. The algorithm allows 6DoF rotation and translation, but for the study only rotation at a fixed listener position was permitted.

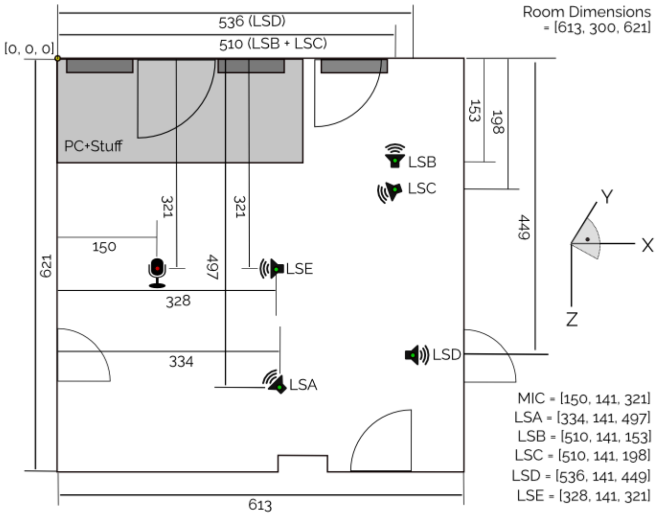

Figure 2: Floor plan of the listening room, including positions of the microphone array and the measured loudspeakers A to E. The listener's seat was located at the former microphone position. Values are in *cm*.

SDM

The Spatial Decomposition Method (SDM) is described in [5, 6]. For measurements it uses one measurement microphone and six electret condensor microphones. It is assumed, that the sound field consists of a sequence of individual acoustic events. They can be described with the captured RIRs and the captured DoAs. In the postprocessing the HRIRs are calculated for the measurement position with 3DoF rotation and a generic HRTF filter.

Listening Test

For the verification of the informal results at the conferences a listening test was conducted. A fully blind indirect comparison of two immersive audio algorithms and real loudspeakers was performed. To do so, a new evaluation method was developed using the HTC Vive Head Mounted Display (HMD) as a tool to show listeners the real room that they were in, without showing the positions of the actual loudspeakers. This avoids a visual bias of the physical loudspeakers on the evaluation. The listeners were first asked to localize the heard sound source inside the room, using the HMD's controller like a laser pointer. Secondly, they were asked to answer whether the heard acoustic events sounded like they were coming from a real loudspeaker inside the room. This was assumed to be a measure for plausibility. Feedback for reasons should be provided, if plausibility was not given.

Measurements and Setup

The dimensions of the measured room are around $6 \times 6\, m^2$ with a volume of circa $114\, m^3$ (figure 2). The room has a symmetric shape, but its acoustics are asymmetric due to a reflective wall with windows in front of the PC.

Five loudspeakers (Genelec 8020D), LSA … LSE, were measured with the SDM measuring array built at TU Ilmenau, either facing towards the microphone position (LSA, LSC and LSE), the windows (LSB) or a door (LSD). The position of the recording equals the position of the listener in the test. The test compared c3k, SDM and the real loudspeakers as a hidden reference (RefLS). A publicly available HRTF dataset of the KEMAR dummy head was applied for binauraulization (Sadie II Database [7]). As headphones STAX SR202 were taken. To get the RIR for c3k only the centered microphone from the SDM array was considered. The HTC Vive Pro HMD was used for a visual representation of the room. It showed a LIDAR (Light Detection and Ranging) scan of the room, captured with an iPad Pro and Polycam app. Preprocessing of the model was done in Blender and rendering in Godot. Figure 3 displays the view the listener was presented through the HMD and the placed indicator of the sound source.

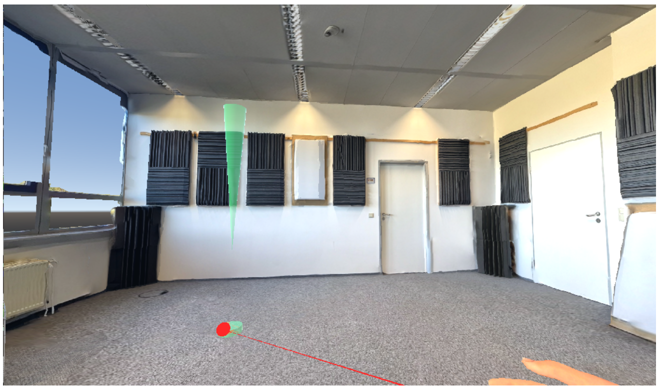

Figure 3: View through the HMD for the listeners including the evaluated position of the sound source.

Conduction

Each listener was led into the room facing away from the test setup. They were placed on a rotating chair inside the room, allowing 3DoF of rotational movement, but no translation inside the room. A female voice recording of part one and five of the Harvard sentence list was used as signal [8, 9].

For training purposes the listeners heard all five loudspeaker positions reproduced by the RefLS, one at a time. They were not aware that they were listening to RefLSs. The test itself included 45 stimuli in total, consisting of 3 methods, 5 loudspeakers and 3 repetitions each.

Results

In total 25 participants took part in the test (17 male, 8 female, average age: 33,5 years – range: 21-68). More than half of them were experienced listeners who had done listening tests on binaural audio before. Most of them were engineers. Others came from completely unrelated professional fields. Four participants were excluded from the analysis. Reasons for that were a misunderstanding of the task, an unsuitable notion of how loudspeakers sound (like a tin can) or conspicuous discontinuities in their ratings.

Figure 4 shows a top view of all estimated positions for all participants. In the left picture the results for the c3k algorithm are shown, in the middle for the RefLSs and on the right for the SDM calculations. It can be seen, that there are different trends in the perception of distance between the representations. In addition to the distance, the perceived angular difference to the original angle of the loudspeaker was calculated.

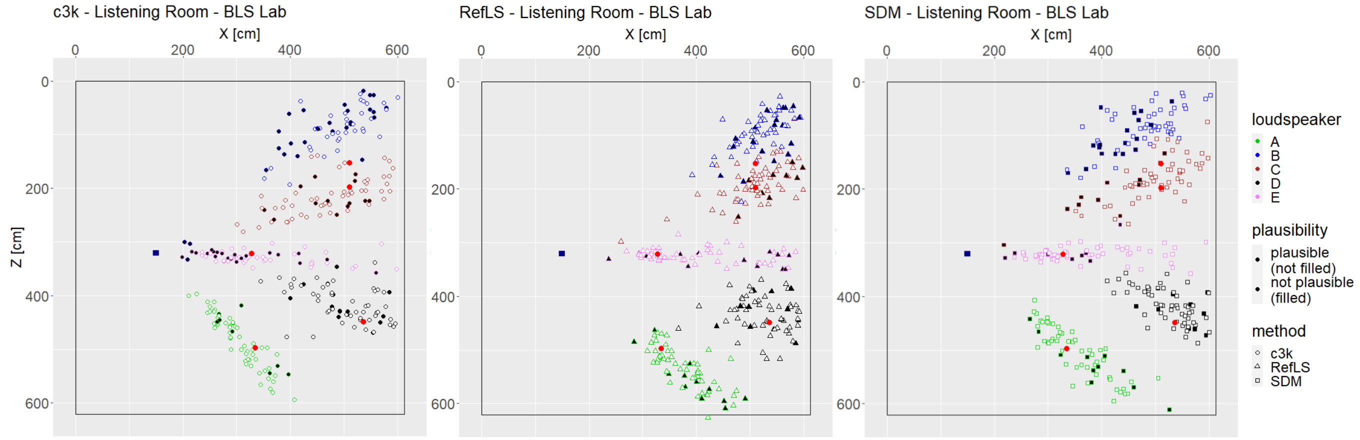

Figure 4: Top view on the room (equals the view from figure 2). The square is the listener position and the red dots are the actual loudspeaker positions. The colors represent the evaluated loudspeakers and the shapes the presented methods. Plausibility is indicated by unfilled shapes.

As an example, results for the distance and angular analysis for LSA and LSB are shown in figure 5. The data is mainly not normal distributed (Shapiro-Wilk Normality Test). Non-parametric test statistics were used to analyze the differences (Kruskal-Wallis Rank Sum Test with post-hoc Pairwise Wilcoxon Rank Sum Test). The tests revealed a few significant differences between some stimuli, but did not show any overall pattern and varied for the different loudspeakers. They did not allow any conclusion and were omitted from this paper. While there were some differences between the medians, standard deviations were quite large and overlapped. Even the RefLSs were not placed at the measured positions. The three representations were rated more similar to each other when faced away from the listener position. Overall tendencies are, that c3k is usually rated nearer, the RefLS mostly a little further away and the SDM closest to the original position. Exceptions were LSB, where all representations tended to be perceived further away, and LSD, where they tended to be closer. Both were not oriented towards the listener position. Loudspeakers faced towards the listener show a smaller angular deviation in the range from −5° to +5°. An extreme case is LSB, since it was oriented towards the windows. The deviation there is in the range from +5° to +15°. As expected, positions oriented away from the listener were perceived offset towards the first reflection and the directly oriented positions were rated around 0°. This shows the importance of the direct sound and the first arriving reflections.

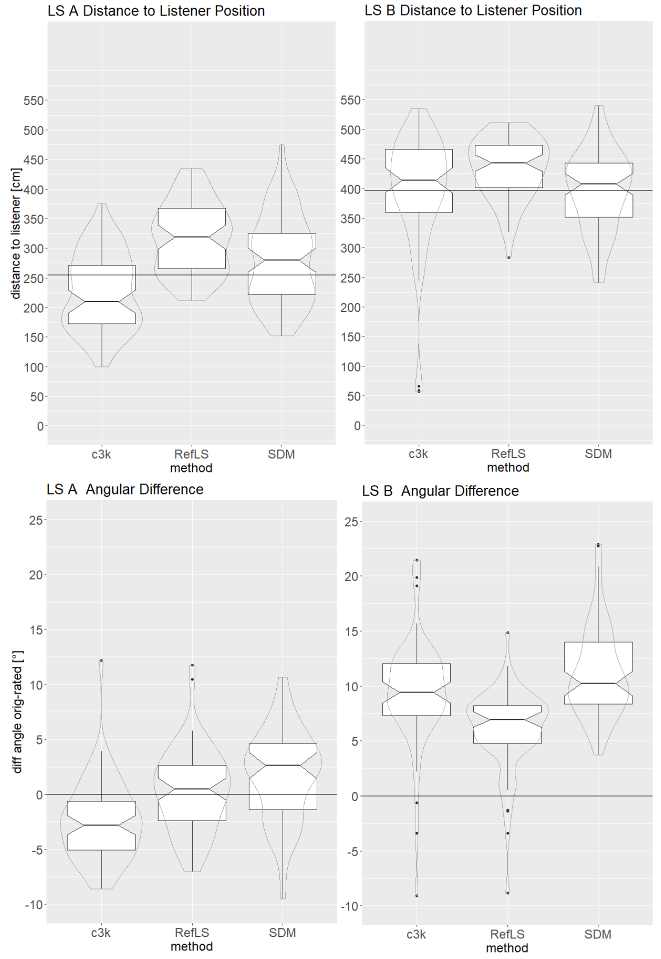

Figure 5: Box plots with violin plots for distance (top) and angle (bottom) for LSA (left) and LSB (right). Solid lines represent distance from listener to loudspeaker position (top) or zero degree angular deviation to the loudspeaker (bottom).

Since the plausibility was evaluated as a “yes” or “no” paradigm a plausibility index was calculated for analysis. The plausibility index is defined as the number of plausible ratings divided by the total number of ratings per condition. The differences for the plausibility index were statistically analyzed using a chi^2 test per loudspeaker. The plausibility was rated similar for all methods, except for LSE, which showed a significant difference between c3k and SDM. The results are shown in figure 6. From the verbal feedback, insights about plausibility impairments could be gained. For c3k it was often stated, that the heard room sounds more diffuse or reverberant, which results in problems regarding timbre perception, perceived source width and distance localization. This happened likely due to an optimization strategy in the algorithm, which reduces the number of rendered early reflections to increase runtime performance. In comparison, the SDM array tended to be rated as plausible more

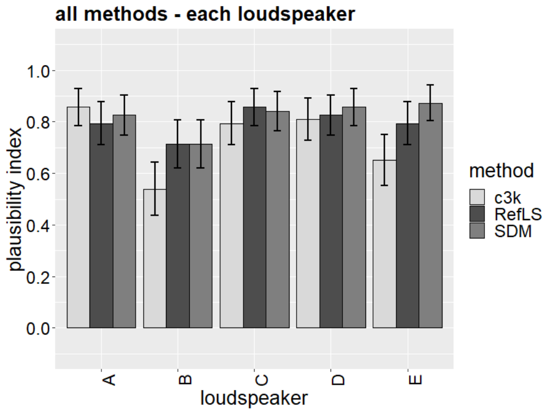

Figure 6: Plausibility index including the standard deviations for all loudspeaker positions and reproduction methods.

often. This was likely caused by the measurement procedure which includes better DoA estimations compared to the one measurement method. The RefLSs were rated as implausible similarly often as the other representations. According to the verbal reports, this might be caused by a perceived instability of the source position. The STAX headphones used in the test are open to the side but have an enclosure to front and back. They slightly filter sounds differently dependent on head rotation. This can possibly be perceived as a positional shift. A differing inner reference of the listening scenario might also be the cause for impairments in plausibility. It is well known that externalization and the mental model of a scene improve, when listeners are allowed to move through the room and explore it. Based on previous tests we assume, that this would increase the plausibility of all three representations to similar, near perfect levels.

Discussion and Outlook

The paper presented a study which was evaluated with a novel method to evaluate and compare real and virtual immersive audio with a fully blind approach. Two binauralization methods, c3k and SDM, as well as playback over real loudspeakers were compared. We found strong deviations between perceived and actual loudspeaker positions regardless of the auralization method. It is well known, that evaluation of localization in distance and angle is difficult without visual cues [10]. The plausibility was rated as similar across all three auralization methods, however a small but not statistically significant impairment of c3k can be suspected. The feedback, especially verbal, is important for further improvements of c3k and other algorithms. It allows us to identify and improve on minor shortcomings of the c3k algorithm. The results further indicate, that both the SDM method and the binaural c3k algorithm are feasible for usage in most real-life applications. Together with the informal feedback on the BLS demo at conferences and exhibitions this proves that a plausible audio illusion with just one measurement is feasible. Multiple interested parties stated, that the qualities of the algorithm would greatly improve their work flow or enable entirely new use cases and products. Current development at BLS focuses on the implementation of c3k for different platforms and use cases. More research on perceived room acoustic differences, efficient rendering of moving sound sources as well as automatic room adjustment procedures is necessary. This and further activities aim to realize the vision of a Personalized Auditory Reality (PARty) [11].

Acknowledgment

The study was part of the two-year joint project TheraTin, which is funded within the framework of the Free State of Thuringia's guideline for the promotion of research, technology and innovation as part of the response of the European Union's response to the COVID-19 pandemic (REACT-EU).

References

[1] 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 8th Int. Conf. on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6.

[2] S. Werner, F. Klein, A. Neidhardt, *et al.*, "Creation of auditory augmented reality using a position-dynamic binaural synthesis system—technical components, psychoacoustic needs, and perceptual evaluation," *Applied Sciences*, vol. 11, no. 3, p. 1150, Jan. 2021.

[3] K. Brandenburg, F. Klein, A. Neidhardt, *et al.*, "Creating auditory illusions with binaural technology," in *The Technology of Binaural Understanding*, Springer, 2020, pp. 623–663.

[4] C. Pörschmann, P. Stade, and J. M. Arend, "Binauralization of omnidirectional room impulse responses - algorithm and technical evaluation," in *DAFx*, 2017.

[5] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *JAES*, vol. 61, no. 1/2, pp. 17–28, Jan. 2013.

[6] L. Treybig, S. Werner, U. Sloma, and G. Stolz, "Measure - analyze - auralize from room impulse response to room classification and binaural reproduction," in *48th Annu. Meet. for Acoustics, DAGA Stuttgart*, 2022.

[7] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database," *Applied Sciences*, vol. 8, no. 11, p. 2029, Oct. 2018.

[8] "IEEE Recommended Practice for Speech Quality Measurements," *IEEE No 297-1969*, pp. 1–24, 1969.

[9] *Miscellaneous anechoic recordings: Harvard Word List*, URL: `https://odeon.dk/downloads/misc-anechoic-recordings/`.

[10] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *ACTA Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005.

[11] K. Brandenburg, E. C. Cerón, F. Klein, *et al.*, "Personalized auditory reality," in *44th Annu. Meet. for Acoustics, DAGA Munich*, 2018.