

The Influence of Visual Cues on the Sound Externalization of Virtual Loudspeaker Setups

CHRISTIAN SCHEER,^{1,*}  FELIX LAU,² CRISTIAN SANDER,²
(christian.scheer@tu-berlin.de) (felix.lau@dear-reality.com) (christian.sander@dear-reality.com)

STEFAN WEINZIERL,¹  AND JOHANNES M. AREND³ 
(stefan.weinzierl@tu-berlin.de) (johannes.arend@aalto.fi)

¹*Audio Communication Group, Technische Universität Berlin, Berlin, Germany*

²*Dear Reality GmbH, Düsseldorf, Germany*

³*Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland*

Research has shown that visual cues influence the perceived sound externalization of binaural reproduction. However, it remains unclear whether providing a virtual visual representation of loudspeakers in addition to their binaural rendering influences their perceived externalization. Using augmented reality (AR), loudspeakers can be visualized within the listener's environment, potentially improving the externalization of binaural loudspeaker renderings. This study presents a listening experiment using dynamic binaural rendering of a typical 7.0.4 immersive loudspeaker setup. Participants rated the externalization of noise bursts rendered as single loudspeaker sources, phantom sources, and moving sources, as well as immersive music, under three different visual conditions. Results show that the AR visualization significantly improves the externalization for noise bursts reproduced by single virtual loudspeakers and for immersive music excerpts reproduced by the entire virtual loudspeaker setup but not for phantom and moving sources. Thus, shifting a percept farther away from the head seems to work best when the acoustically perceived sound direction matches the visual source location. The effectiveness of AR visualization for immersive music, on the other hand, may be due to listeners' familiarity with spatial music productions over loudspeakers.

Keywords: sound externalization, binaural audio, spatial audio, virtual acoustic realities, virtual loudspeaker setups, 6DoF, augmented reality, mixed reality, auditory distance perception, psychoacoustics

1 INTRODUCTION

Immersive audio and 3D audio technology continues to grow in popularity in both consumer and professional environments, with consumer products such as Apple AirPods and Apple Spatial Music and with Dolby Atmos and MPEG-H production tools in the professional context. To produce immersive audio content, current 3D audio standards recommend the use of multiloudspeaker systems with surrounding loudspeakers at different heights in acoustically optimized control rooms [1–4]. However, such a setup is not always affordable, which is why a number of software solutions for binaural loudspeaker rendering and headphone-based reproduction of virtual loudspeaker se-

tups have emerged. These tools, including the proprietary renderers of the aforementioned 3D audio standards as well as audio plug-ins such as dearVR MONITOR, APL Virtuoso, and Noisemakers Binauralizer Studio 2 provide an accessible and portable method for simulating common immersive audio loudspeaker setups in different rooms [5–7].

1.1 Binaural Rendering

Binaural rendering of a virtual loudspeaker array in a room is typically based on a set of binaural room impulse responses (BRIRs), where each BRIR describes the transfer path from a particular loudspeaker to the listener. In its simplest form, rendering is performed by convolving the output of each virtual loudspeaker with its respective BRIR, and summing the resulting signals channel-wise to produce a two-channel binaural signal that is then delivered to the listener via headphones. For binaural rendering with three degrees of freedom (3DoF), where listeners can rotate their head in yaw, pitch, and roll, BRIRs are required for each head orientation. For rendering with 6DoF, where listeners are also free to move around the room, BRIRs are needed for every combination of listener position and head

*To whom correspondence should be addressed, email: christian.scheer@tu-berlin.de

orientation. For any combination of such dynamic binaural renderings, the listener is tracked and the convolution pipeline dynamically switches between BRIRs corresponding to the listener's current position and/or head orientation to update the rendering.

The required BRIRs can be measured either with a dummy head or individually [8–10], synthesized based on microphone (array) measurements [11–14] or solely based on a given parametric sound field description [15–17], or simulated using room acoustic simulation engines [18, 19]. The different techniques are often combined into hybrid methods so that, for example, the direct sound and early reflections are simulated, but the binaural reverberation comes from a measurement or parametric synthesis.

Measuring BRIRs is the most time-consuming and least flexible of these approaches, especially when dealing with multichannel setups, as artifact-free 3DoF binaural rendering would require measurements of each loudspeaker with a 2° resolution for yaw and pitch head orientations [20], and 6DoF rendering would additionally require these measurements to be made with a positional resolution as fine as 1 m with interpolation and 5 cm without [21, 22]. The measurement effort can be reduced by measuring BRIRs for only a few head orientations and positions in the listening space and generating the remaining required BRIRs using (non)parametric interpolation methods [23, 24]. However, reducing the number of measurements, even when using advanced interpolation methods, may result in a loss of reproduction quality with audible artifacts.

BRIR synthesis based on microphone (array) measurements further reduces the measurements effort, as one measurement per loudspeaker is sufficient for 3DoF rendering [25], and various interpolation and extrapolation approaches exist to reduce the measurement effort when aiming for 6DoF-rendering [14, 26, 27]. For BRIR synthesis, measurements can be processed using nonparametric methods—typically Ambisonics encoding and binaural decoding [28]—or parametric methods, where the sound field is decoded into parameters used to generate the BRIRs [29] (see also [30] for an overview). In addition, these approaches are more flexible in terms of listener individualization, as any set of head-related transfer functions (HRTFs) can be used for the synthesis.

Finally, room acoustic simulation or hybrids between simulation and parametric synthesis are the most flexible approaches, allowing real-time computation of individualized BRIRs for any listener's head orientation and position and, most importantly, being able to quickly adapt to changes in the loudspeaker setup or the virtual listening room with its room acoustic properties. Notably, some recent real-time room acoustic simulation engines do not compute full-length BRIRs for convolution-based rendering, which requires significant computational effort. Instead, these engines process incoming audio buffers using cascaded filter-delay structures to simulate direct sound and early reflections according to an image source model (ISM) and reverberation using feedback delay networks and spatialize the ISM and feedback delay network signals by convolution with short HRTF filters, making the sim-

ulation engine much more computationally efficient [31, 32]. However, the effort required to implement efficient real-time simulation algorithms that provide perceptually plausible binaural rendering remains high.

1.2 Externalization

Sound externalization, that is, the perception that a sound source is outside of the head [33], is a key perceptual aspect of binaural rendering. Externalization is particularly important for binaural rendering of virtual loudspeaker setups, where the goal is for sound engineers to experience the sound over headphones as if they were in a control room with a loudspeaker array surrounding them. From the earliest studies on externalization, researchers have found that achieving externalized sound images requires the reproduction of “ear-adequate” signals, meaning that the binaural rendering should contain all the cues of a natural external sound, such as source, room, and listener characteristics [34]. Subsequent studies then identified several factors that influence perceived externalization, which can be broadly categorized into acoustic cues in direct sound and reverberation, as well as multisensory factors [33].

Regarding direct sound cues, researchers have found that lateral virtual sound sources, which provide strong binaural difference cues in the form of interaural time and level differences, tend to be better externalized than median plane sources, where interaural differences are small [35, 36]. Studies on the impact of individual HRTFs on perceived externalization are inconclusive, with some results suggesting no effect [37, 38] and others suggesting an effect [39, 40] of individualization, with individualized renderings leading to better externalization especially for median-plane sources in the latter study. Summarizing the influence of individual HRTFs on perceived externalization, Best et al. concluded in their review that individual HRTFs may not be critical for externalization in more realistic listening situations that include reverberation and multisensory cues [33].

When the virtual sound source contains reverberation in addition to the anechoic direct sound, it is generally more likely to be perceived as externalized [37, 38]. Interestingly, early reflections up to 80 ms after the direct sound are sufficient to improve externalization, whereas including the entire reverberation tail in the rendering does not further improve externalization [37, 41]. Moreover, the spectral detail in the reverberation is less important for externalization than that in the direct sound [42]. Rather, sufficient binaural information in the reverberant part is critical for externalization, such as binaural cues from lateral reflections or fluctuations in interaural level difference and related interaural coherence (IC) [43, 41, 38]. Further investigation of these fluctuations revealed that for lateral sound sources, especially the reverberation at the contralateral ear showing strong fluctuations is critical for externalization [44].

In terms of multisensory factors, head movements and vision are particularly relevant to externalization. Head movements do not only improve localization accuracy and help resolve front-back confusion [45] but also contribute significantly to externalization perception [36, 46], especially

for median-plane sources for which head movements induce binaural difference cues [33]. Interestingly, the effect is stronger for self-motion than for source movements [47], and externalization can persist even when the listener is no longer moving [46].

In a multisensory scene, however, neither acoustic cues nor head movements may be robust enough to provide an externalized sound image. Rather, the binaural rendering must also be consistent with the listener's expectation of the real room and its acoustic properties, as assessed visually and/or auditorily, in order to be perceived as externalized. If the binaural rendering is not congruent with the listener's expectation of how a source should sound in a particular room, externalization may break down and the listener may perceive the virtual sound source as (more) internalized, which is called the room divergence effect [40, 48–50]. Thus, a visual impression of the room that is congruent with the binaural signals facilitates or even improves externalization [40, 48, 49].

1.3 Aim of this Study

In summary, previous research has revealed important acoustic and multisensory factors that influence perceived externalization in binaural rendering, with visual cues of the listening room congruent with the auditory scene being a key factor for externalized perception. However, there is no systematic and detailed research on how the visual presence of a loudspeaker at the position of the virtual sound source affects perceived externalization, and in particular, there is no research on such possible effects in the case of a *virtual visual representation* of the loudspeaker [i.e., when the loudspeaker is augmented in the real listening room using augmented reality (AR) technology].

Furthermore, previous externalization studies have mostly considered discrete sound sources emitted by a single loudspeaker but not static phantom sources, moving sources, or even entire immersive audio productions, which, however, is crucial for binaural reproduction of immersive audio setups. It therefore remains an open research question whether the visual augmentation of a loudspeaker or even a loudspeaker array surrounding the listener, thereby creating a stronger sense of being in an immersive audio setup, affects and potentially enhances the perceived externalization of the binaural reproduction of such a setup.

This paper addresses these questions by systematically investigating how different degrees of loudspeaker visualization affect the perceived externalization of a parametric 3DoF+ (see SEC. 2.3) binaural rendering of a 7.0.4 immersive audio setup. To this end, a listening experiment was conducted in which participants rated the perceived externalization of noise bursts rendered as single loudspeaker sources, phantom sources, and moving sources, as well as immersive music, under three visual conditions: (1) AR visualization of the surrounding loudspeaker setup using a mixed-reality headset, (2) 2D visualization of the loudspeaker setup on a conventional computer screen, and (3) no additional visual information.

2 METHOD

2.1 Participants

A total of 33 participants (age 23–43, $M = 29.1$ years, $Mdn = 28$ years, $SD = 5.02$) with self-reported normal hearing participated in the experiment. Most of them can be considered experienced listeners, with nine of them being members of the Audio Communication Group or working for audio companies and 17 reporting active contact with acoustics or music in their professional or personal lives. All participants were naive as to the purpose of the study, and all gave informed consent for participation in the study and publication of the results.

2.2 Setup

The listening experiment was conducted in a control room ($L = 9.2$ m, $W = 7.8$ m, $H = 3.3$ m, $RT_{1\text{ kHz}} = 0.34$ s, $L_{Aeq} = 20$ dB background noise) at TU Berlin, which reflects a typical immersive audio production environment. To avoid any potential visual distraction for the participants, most of the permanently installed loudspeakers in the listener's line of sight that are normally installed in the room were removed. The permanently installed ceiling loudspeakers that could not be removed were covered with visually opaque textiles. Overall, this resulted in an appropriate setup for the study without critical visual distractions [see Fig. 2(a)].

Fig. 1 illustrates the technical setup used in the listening experiment. A custom Python application running on a PC controlled the test procedure as well as the binaural rendering software using Open Sound Control commands. The graphical user interface (GUI) of the control application was displayed on a large screen (diagonal width of about 80 in and a distance to the listener of about 2 m) to ensure that it was still usable while wearing the mixed-reality headset.

An RME Fireface UFX+ audio interface running at a sampling rate of 48 kHz and a buffer size of 512 samples performed the digital-to-analog conversion. The analog headphone output of the Fireface was connected to a Sennheiser EW IEM G4 E-Band wireless monitoring system, which fed Sennheiser HD 490 Pro headphones, allowing the listener to move freely around the room.

A HTC VIVE Tracker 3.0 was mounted on top of the headphones and tracked by four HTC VIVE SteamVR Base Stations 2.0. The head orientation and position data were processed with a custom Unreal Engine 5.1 application and then sent via Open Sound Control at an update rate of 100 Hz to the binaural rendering software, which was implemented in Pure Data.

The 2D visualization of the loudspeaker setup [see Fig. 2(b)] was displayed on a 19 in computer screen placed directly in front of the subject, while maintaining a full view of the large screen in the back. For the AR visualization of the surrounding loudspeaker setup, an Apple Vision Pro mixed-reality headset was used in combination with a custom standalone visionOS application implemented in Swift and with Apple's own inside-out tracking mechanism. Fig. 2(c) shows a screenshot of the image displayed in the

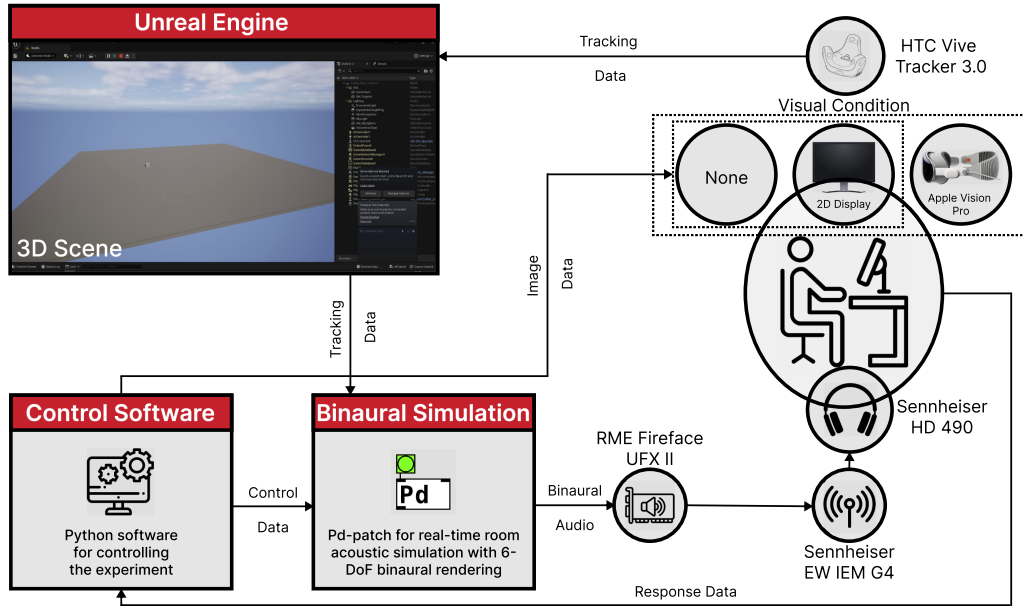


Fig. 1. Overview of the technical setup used for the listening experiment.

Apple Vision Pro, with the virtual visual representation of the loudspeakers augmented in the real room. Note that the same visual scene of the complete 7.0.4 loudspeaker setup was shown for both the 2D and AR visualization, and there were no visual cues indicating which loudspeakers were currently active.

2.3 Binaural Rendering Engine

For the room acoustic simulation and binaural rendering of the virtual 7.0.4 loudspeaker setup, an ISM in a rectangular (shoebox) room for real-time simulation and listener-dependent adaptation of direct sound and early reflections was combined with a static diffuse BRIR generated using frequency-dependent decaying noise for reverberation. Fig. 3 shows the signal flow diagram of the Pure Data implementation. As the diagram shows, the renderer accepts a mono input signal for each channel of the loudspeaker configuration, processes the audio buffers of these channels in real time to generate the loudspeaker and room acoustic simulation, and provides a two-channel binaural signal at the output.

The implementation uses several finite impulse response filters for room acoustic simulation and binaural rendering. The present study employed a dense full-spherical set of Neumann KU100 HRTFs for spatialization, dense full-spherical directivity measurements of the simulated Neumann KH120 II to incorporate loudspeaker directivity, and a headphone compensation filter for the combination of Neumann KU100 and Sennheiser HD 490 Pro. Further, the simulation employs custom wall reflection filters obtained for a collection of materials based on absorption coefficients from the room acoustic absorption coefficient database of the Physikalisch-Technische Bundesanstalt [51]. For matching the simulation to the acoustics of the control room (see SEC. 2.4), appropriate materials were selected from this filter database.

Table 1. Virtual loudspeaker positions in azimuth ϕ and elevation θ around the sweet spot.

No.	Name	ϕ	θ
1	Left (L)	30°	0°
2	Right (R)	-30°	0°
3	Center (C)	0°	0°
4	Rear Left (RL)	135°	0°
5	Rear Right (RR)	-135°	0°
6	Surround Left (SL)	90°	0°
7	Surround Right (SR)	-90°	0°
8	Front Up Left (FUL)	45°	45°
9	Front Up Right (FUR)	-45°	45°
10	Rear Up Left (RUL)	135°	45°
11	Rear Up Right (RUR)	-135°	45°

Note. The radius for all loudspeakers was $r = 1.50$ m.

The simulation uses the closed-form ISM for efficient direct sound and early reflection calculation [52, 53]. Based on the room dimensions (in the present case, $L = 9.23$ m, $W = 7.75$ m, $H = 3.32$ m), the dynamic listener position, and the static loudspeaker positions in the room, the ISM implementation constantly calculates the time and direction of arrival of the direct sound and the image sources for each loudspeaker. Table 1 shows the loudspeaker positions with respect to the initial listener position (i.e., the sweet spot), which for this study was defined as $X = 5.73$ m, $Y = 4.01$ m, and Z being the individual height of the participant's interaural axis.

For direct sound rendering (see upper path per input in Fig. 3), the audio buffers of each input channel enter the *Direct* renderer, where they are first delayed using a variable delay implementation and gain-adjusted according to the current distance between the loudspeaker and the listener. Next, the renderer selects the correct HRTF according to the current direction of arrival with respect to the listener's position and head orientation, as well as the correct source

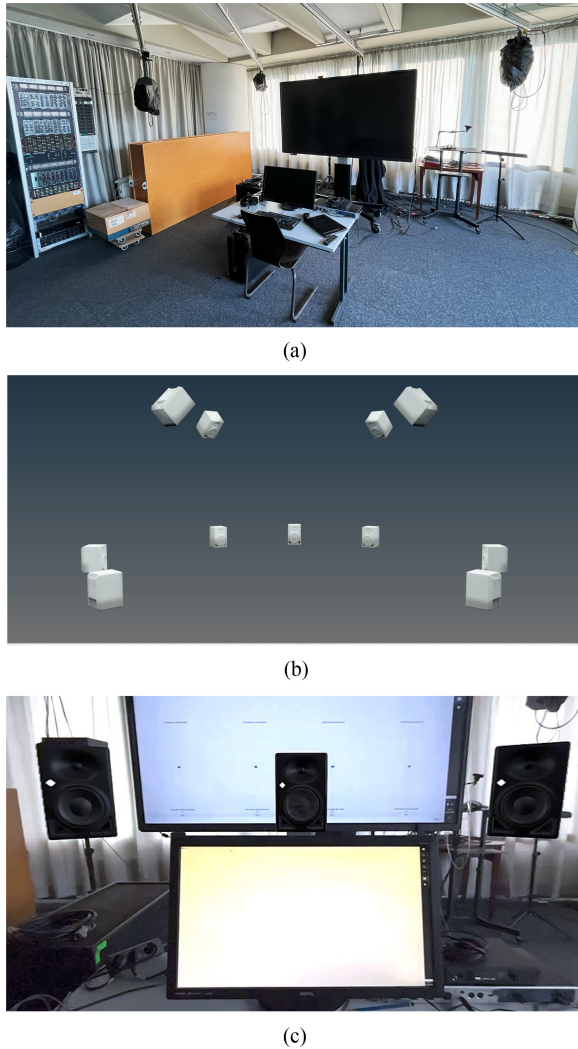


Fig. 2. Experimental setup showing a large screen at the back for displaying the experiment software's GUI and a smaller screen in front of the listener for presenting the 2D visualization of the virtual loudspeaker setup (a). 2D visualization of the virtual 7.0.4 setup, consisting of seven loudspeakers in a horizontal ring and four in an elevated ring. Table 1 lists the exact position of each loudspeaker (b). Screenshot of the view inside the Apple Vision Pro, currently displaying three frontal loudspeakers of the augmented 7.0.4 virtual loudspeaker setup in the real room (c).

directivity filter according to the current source exit angle, and combines both filters by fast convolution to one direct sound filter. Finally, this two-channel direct sound filter is applied to the audio buffers using uniformly partitioned convolution.

For the early reflections (see lower path per input in Fig. 3), the input signals first pass through a global reflection delay and gain, which can be used to adjust the overall time of arrival (similar to a pre-delay parameter) and level of the early reflections. In the following *Reflections* renderer, the positions of the image sources are calculated based on the positions of the loudspeakers and the room dimensions. In this study, only the first-order image sources were calculated, resulting in six image sources per loudspeaker. This was deemed sufficient because the first order typically encompasses the most salient early reflections [54, 55], and

using only the first order makes the simulation less computationally demanding.

The image source positions are calculated once and updated only when the loudspeaker positions change. For each image source, a two-channel filter is generated as follows. The renderer selects an HRTF for the direction of the listener's current position and head orientation relative to the position of the image source, delays and gain-adjusts this HRTF according to the current distance between the image source and the listener, filters this HRTF with the respective wall reflection filter, and lastly applies the source directivity filter based on the current source exit angle to this processed HRTF. Next, the left and right channels of the six image sources are summed, resulting in a single two-channel early reflections filter per loudspeaker, which is applied to the input signal using nonuniformly partitioned convolution.

For each loudspeaker, the direct sound and early reflection signals are summed channel-wise, resulting in a two-channel signal per virtual loudspeaker. Finally, the left and right channels across all virtual loudspeaker signals are summed separately, yielding a two-channel binaural signal that contains the direct sound and early reflection components from all virtual loudspeakers corresponding to the user's current head orientation and position in the room.

To render the static late reverberation, all input channels are summed first (see lower part of Fig. 3). The resulting mono signal can be delayed and level-adjusted using global reverberation delay and gain parameters, and is then convolved with a diffuse BRIR using non-uniformly partitioned convolution. For this study, the diffuse BRIR was generated using the MATLAB function `AKdiffuseReverbTail` from `AKtools` [56], with the control room's reverberation times T_{20} in 1/3-octave bands as the input parameters. The function uses the target T_{20} to generate a binaural diffuse reverberation tail from frequency-dependent decaying noise, including binaural coherence matching [57]. Additionally, a fade-in was performed toward the mixing time on the late reverberation tail [58]. The resulting two-channel binaural late reverberation signal is summed together with the direct sound and early reflection components, providing the full binaural simulation of the virtual loudspeaker setup. Finally, the binaural signal is processed with the headphone compensation filter and sent to the stereo output that feeds the user's headphones.

It is worth noting that in the present implementation of the renderer, the distance-dependent gain adjustments for the direct sound and the early reflections did not strictly follow the inverse square law. Based on informal listening tests, the distance-dependent gain adjustments were decreased to reduce the influence of the listener's transitional motion on the loudspeaker and image source levels. This greatly stabilizes the level balance at the sweet spot when the listener moves their upper body, thereby improving the overall quality of the reproduction. In turn, changes in 6DoF rendering are largely reduced to translation-dependent adjustments in the localization cue, which, in agreement with Amengual Garí et al. [50], is defined here as 3DoF+ rendering. However, the authors believe that this modification had no effect on the experiment because the participants were instructed

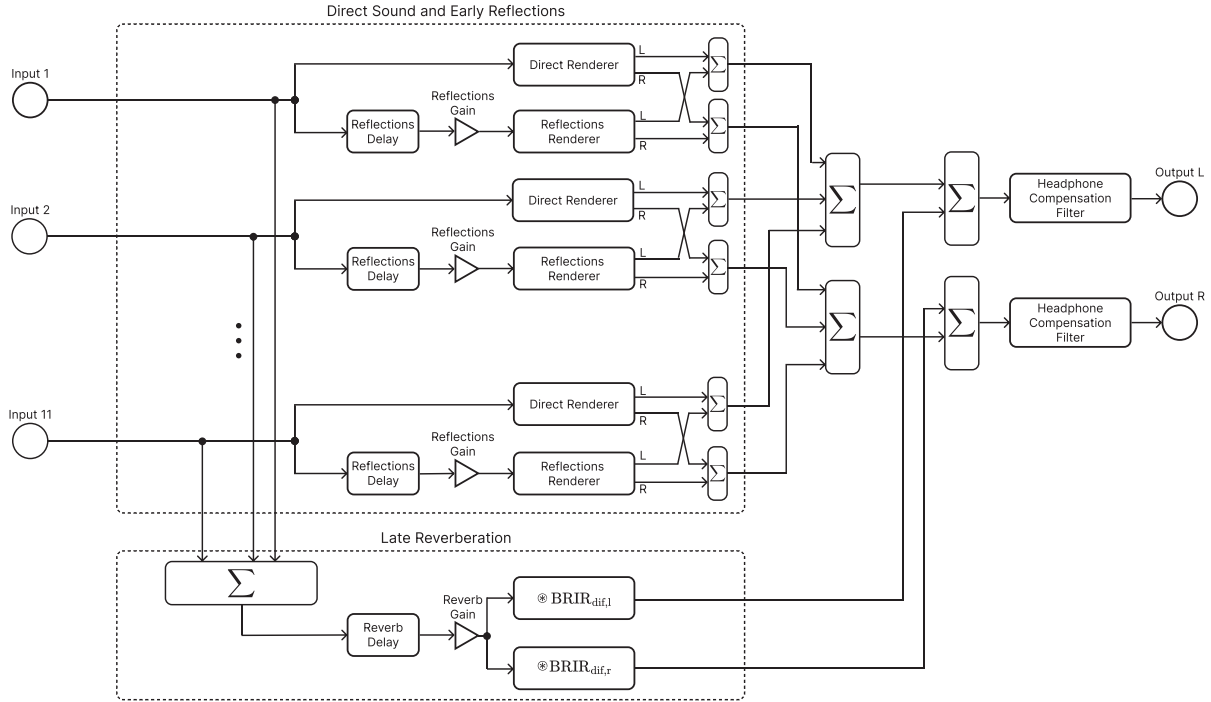


Fig. 3. Signal flow diagram of the binaural 7.0.4 loudspeaker and room acoustic simulation.

to remain seated in the sweet spot and, thus, there were no noteworthy changes in distance to the virtual loudspeakers and their image sources.

2.4 Room Matching

To avoid room divergence effects, the simulated acoustics were matched to the real acoustics of the control room through an iterative process in which the simulation was adjusted according to acoustic parameters of the real room. For this, a reference BRIR was measured with the swept sine method using a real Neumann KH 120 II at the center loudspeaker position of the 7.0.4 setup and a Neumann KU81i dummy head at the sweet spot, facing the loudspeaker with frontal head orientation.

This BRIR was used as a reference to estimate perceptually relevant room acoustic parameters, which served as target values for the simulation. These included the broadband parameters reverberation time ($T_{20} = 0.364$ s), direct-to-reverberant ratio ($DRR = 8.15$ dB), and interaural cross-correlation ($IACC = 0.772$). Additionally, the average reverberation time of the control room was estimated in 1/3-octave bands. This estimation was based on 15 measurements with an omnidirectional microphone (NTI M2230) for three loudspeaker positions and five microphone positions for each loudspeaker position. The determined frequency-dependent T_{20} values of the control room served as input parameters for synthesizing the initial late reverberation tail, as described above in SEC. 2.3.

To obtain the simulated BRIR, the binaural rendering engine was measured through the center channel using the swept sine method. In an iterative process of comparing the measured and simulated BRIR, the late reverberation

tail was adjusted by fine-tuning the T_{20} input parameters of the synthesis, the fade-in toward the mixing time, and the renderer's reverberation gain parameter. Additionally, during this process, the ISM parameters were thoroughly adjusted, selecting suitable wall reflection filters from the pregenerated material filters and adjusting the overall gain of the early reflections to align their energy levels in the simulated and measured BRIRs. The aim of this iterative process was that the entire simulated BRIR (including direct sound, early reflections, and late reverberation) would match the measured BRIR in terms of the broadband parameters T_{20} , DRR, and IACC.

With the final simulation settings, the measured and simulated BRIR showed an absolute DRR difference of 0.89 dB, which is well below the just-noticeable difference (JND) of 2 dB [59]. Also, the absolute IACC difference of 0.052 was below the JND of 0.075 (according to ISO 3382-1), as was the relative difference in broadband T_{20} , which was 4.65% and thus remained below the 5% JND threshold [60].

Fig. 4 compares the measured and simulated BRIRs. The broadband pressure BRIRs for the left ear [Fig. 4(a)] show good similarity in their overall time-energy structure with matching amplitude and time events. Similarly, the left ear binaural room transfer functions (BRTFs) in Fig. 4(b) show good agreement, although the analysis reveals magnitude differences of up to 5 dB at higher frequencies (disregarding the low-frequency roll-off in the measured BRIR), most likely due to the fact that the measurement was performed with a Neumann KU81i dummy head while the simulation used Neumann KU100 HRTFs.

The reverberation times T_{20} of the simulated and measured BRIR also show good agreement [Fig. 4(d)]. The

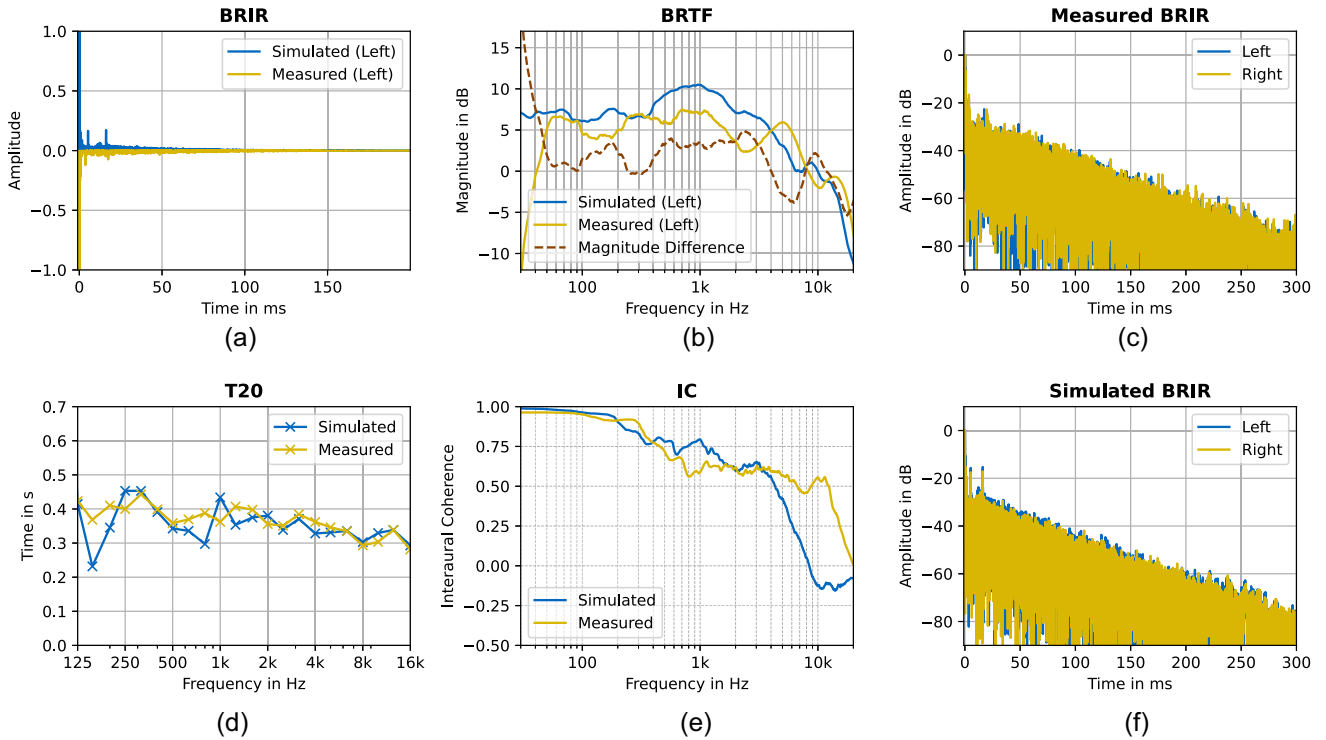


Fig. 4. Comparison of measured and simulated BRIRs of the center loudspeaker at the sweet spot in terms of time-energy structure (a, c, and f), magnitude response (b), reverberation time averaged over left and right ear with error bars indicating the interaural standard deviation (d), and IC (e). Note that (a) shows absolute values, and that the measured BRIR (yellow) has been inverted for easier visual comparison.

deviation in the lower frequencies (especially at 160 Hz) is most likely due to the matching of the diffuse BRIR synthesis to the broadband T20 of the measurement and not to the T20 in 1/3-octave bands.

The direct comparison of the frequency-dependent IC in Fig. 4(e) shows a good agreement up to about 3 kHz, with a generally rather high IC value above 0.6 for both BRIRs, mostly because the sound source was in the front and the room is rather dry, and without strong (uni)lateral reflections, leading to similar ear signals. Above 3 kHz, the IC of the simulated BRIR decreases rapidly and deviates from the measurement. This is most likely because the simulated BRIR contains a reverberation tail with binaural coherence for an ideal diffuse field, which is not fully satisfied in the real room.

Finally, Figs. 4(c) and 4(f) illustrate both BRIRs on a logarithmic amplitude scale, highlighting the similar decay and energy structure of both BRIRs.

2.5 Stimuli

Table 2 shows the experimental conditions for which subjects had to rate perceived externalization. The experiment assessed externalization depending on the factors visualization (AR, Display, None) and source type (Single, Phantom, Moving, Music), with different numbers of variants depending on the source type (four single sources, four phantom sources, three moving sources, two music pieces).

Single, phantom, and moving sources were all different spatial variations of the same 7.7 s long sequence of 100 ms pink noise bursts (including 2-ms cosine-squared

onset/offset ramps) with an interstimulus interval of 300 ms. Pink noise was chosen because of its (well-describable) broadband spectrum and because previous studies have provided some evidence that pink noise is perceived as more externalized than, for example, speech [38].

Regarding the different source type conditions, single sources describe stimuli where only one specific virtual loudspeaker emitted the pulsed noise, phantom sources describe stimuli that were statically panned to a position between the virtual loudspeakers using vector-base amplitude panning (VBAP) [61], and moving sources describe stimuli that moved along prescribed trajectories within the loudspeaker setup using VBAP. VBAP was used because it is a widely adopted, robust, practical approach for amplitude panning in immersive loudspeaker setups [28]. Table 3 summarizes the different variants per source type. The presentation level for all noise stimuli was $L_{Aeq} = 54.2$ dB, measured as free-field sound pressure level next to the ear. The supplementary material [62] provides static binaural renderings of the different noise stimuli to give the reader an impression of the reproduced sound.

For the music condition, two 90 s excerpts of 7.1.4 Dolby Atmos pop mixes (“Rocket Man” by Elton John and “I Didn’t Change My Number” by Billie Eilish) were used. These were chosen because they are high quality productions with rather different mixing approaches. The “Rocket Man” mix follows a more conventional approach to 3D audio mixing, with the main vocal track placed mainly on the center loudspeaker and the surround channels used primarily for reverberation. In contrast, the “I Didn’t Change

Table 2. Experimental conditions for which subjects had to rate perceived externalization, with different numbers of variants depending on the source type.

Visualization	AR				Display				None			
Source type	Single	Phantom	Moving	Music	Single	Phantom	Moving	Music	Single	Phantom	Moving	Music
No. of variants	4	4	3	2	4	4	3	2	4	4	3	2

My Number” mix makes minimal use of the center loudspeaker, leaving it free of the main vocals. In addition, the surround channels are used not only for reverberation but also for subtle effects, such as the dog sounds coming from the back at the beginning. To reproduce the 7.1.4 mixes with the 7.0.4 loudspeaker simulation, the LFE channel was routed to the center channel of the virtual loudspeaker system (input 3 in Fig. 3). The presentation level for both (loudness equalized) music examples was $L_{A_{eq}} = 72.2$ dB, also measured as free-field sound pressure level next to the ear.

2.6 Procedure

The experiment was designed as a within-subjects study, meaning that each participant rated the perceived externalization of all conditions listed in Table 2. Participants were asked to rate the perceived externalization on a scale from 0 (perceived inside the head) to 1 (completely externalized), following the rating paradigm proposed by Leclère et al. [38]. The study followed a randomized block design, i.e., the experiment was divided into three blocks, presented in randomized order, each containing all acoustic conditions per visualization condition (AR, Display, None). Each block was then divided into four sub-blocks, also presented in randomized order, each containing all source type variants for one source type (Single, Phantom, Moving, and Music). Within each sub-block, the order of source type variants was again randomized.

Participants were given detailed instructions about the test procedure and the concept of externalization, although they were not informed about the aims of the study. Prior to each experimental session, participants completed a training session to familiarize themselves with the test procedure and experimental design. The examples used for training were: a single source from the center loudspeaker, a phantom source at $(\phi, \theta) = (45^\circ, 0^\circ)$, a source moving from back to front across the left side [i.e., $(\phi, \theta) = (180^\circ - 0^\circ, 0^\circ)$], and the excerpt of “Rocket Man” as a music example.

Participants rated externalization on a scale of 0 to 1 by moving a continuous slider with extremities labeled “per-

ceived inside the head” and “completely externalized” displayed on the GUI with a computer mouse. Each rating page per sub-block contained all variants for the current source type, and participants could listen to each variant as many times as they wanted. After rating all variants for the current source type, clicking the “Next” button took participants to the next rating page corresponding to the next source type. They could not return to the previous page.

During the experiment, participants were instructed to remain seated in the sweet spot, but to make typical head movements as they would when working with audio material in an immersive loudspeaker setup. The videos in the supplementary material illustrate the experimental procedure and the different visualization conditions [62].

2.7 Data Analysis

For the statistical analysis of the externalization scores, a total of six different aggregations or subsets of the data (further detailed in the results, SEC. 3) were created. Visual inspection of these and Shapiro-Wilk tests for normality, corrected for multiple hypothesis testing according to Hochberg [63], revealed slight violations of normality for only one condition in the moving source subset (condition visualization AR, trajectory horizontal right). Given the robustness of parametric testing to slight violations of normality assumptions [64], the data was analyzed using Greenhouse-Geisser (GG)-corrected [65] repeated measures analyses of variance (ANOVAs). All ANOVA post hoc tests were Holm-corrected to adjust the initial significance level of 0.05 for multiple comparisons. Single pairwise-comparisons using t tests were Hochberg-corrected. Data analysis was done with R, using the `ezANOVA` package for the repeated measures ANOVAs and the `superb` package [66] for calculating the within-subject confidence intervals [67].

3 RESULTS

For an overall analysis of the effect of *visualization* on externalization, aggregated results for visualization were

Table 3. The different variants for the four source types.

Single	Phantom	Moving	Music
Loudspeaker	Position (ϕ, θ)	Trajectory (ϕ, θ)	Song
Center (C)	$45^\circ, 0^\circ$	$180^\circ - 0^\circ, 0^\circ$ (Horizontal left)	Elton John
Left (L)	$115^\circ, 0^\circ$	$0^\circ, 0^\circ - 45^\circ$ (Vertical front)	Billie Eilish
Front Up Left (FUL)	$45^\circ, 20^\circ$	-180° to $0^\circ, 0^\circ$ (Horizontal right)	
Rear Left (RL)	$115^\circ, 20^\circ$		

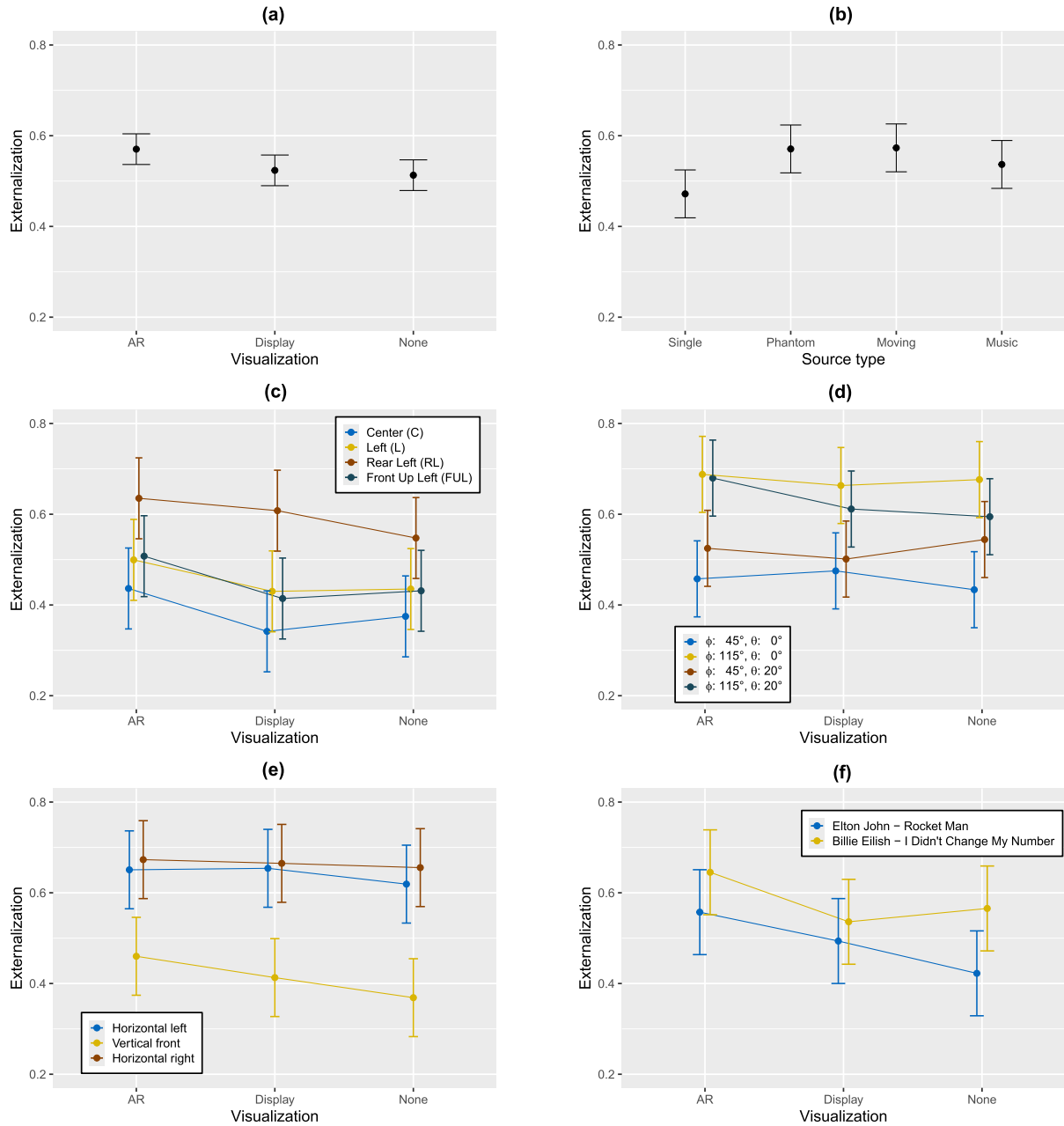


Fig. 5. Results of the listening experiment as mean plots for the aggregated results (a, b) and the source type subsets (c, d, e, f). The aggregated results for visualization (aggregated means) and source type (aggregated means) were obtained by averaging each participant's externalization ratings across source type and visualization, respectively. The error bars indicate difference and correlation-adjusted (Loftus-Masson method) 95% within-subject confidence intervals [66, 67]. (a) Visualization (aggregated means); (b) source type (aggregated means); (c) single; (d) phantom; (e) moving; (f) music.

generated by averaging each participant's externalization ratings across source types (i.e., 13 ratings per visualization condition), resulting in one aggregated mean externalization rating per participant for each visualization condition. The results plot in Fig. 5 suggests that the AR condition yielded slightly higher externalization ratings than the other conditions, and the corresponding ANOVA revealed a significant effect of visualization (see Table 4). Post hoc comparisons confirmed that the AR visualization led to significantly higher externalization ratings than the display visualization [$t(32) = 3.14$, $p = 0.011$, 0.57 ± 0.13 vs. 0.52 ± 0.15] or no visualization [$t(32) = 3.15$, $p = 0.011$,

0.57 ± 0.13 vs. 0.51 ± 0.14], whereas no significant difference was found between display and no visualization [$t(32) = 0.64$, $p = 0.53$, 0.52 ± 0.15 vs. 0.51 ± 0.14]. However, as the following analysis will show, the effect of AR visualization is not consistent across source types and is mainly driven by the single and music source type conditions.

To provide a more comprehensive view of the effect of *source type* on externalization, aggregated results were also generated for source type by averaging each participant's externalization rating across visualizations (i.e., 12 ratings for single and phantom, nine ratings for moving, and six

Table 4. Results of the six repeated measures ANOVAs to analyze the aggregated results for visualization and source type, as well as the source type subsets. The aggregated results for visualization (aggregated means) and source type (aggregated means) were obtained by averaging each participant's externalization ratings across source type and visualization, respectively. The source type subsets contain the corresponding raw data. The different source type variants are listed in Table 3.

Dataset	<i>df</i>	<i>F</i>	MSE	ϵ	η_p^2	<i>p</i> _{GG}
Visualization (aggregated means)						
Visualization	2, 64	6.76	0.005	0.95	0.18	0.003**
Source type (aggregated means)						
Source type	3, 96	6.66	0.019	0.60	0.17	0.003**
Single						
Visualization	2, 64	4.92	0.048	0.96	0.13	0.011*
Loudspeaker	3, 96	16.36	0.058	0.83	0.34	<0.001***
Visualization $\times$ loudspeaker	6, 192	0.84	0.028	0.65	0.03	0.501
Phantom						
Visualization	2, 64	0.81	0.044	0.78	0.03	0.425
Position	3, 96	25.14	0.052	0.76	0.44	<0.001***
Visualization $\times$ position	6, 192	1.23	0.032	0.63	0.04	0.304
Moving						
Visualization	2, 64	1.43	0.039	0.98	0.04	0.247
Trajectory	2, 64	51.72	0.057	0.65	0.62	<0.001***
Visualization $\times$ trajectory	4, 128	0.67	0.026	0.80	0.02	0.583
Music						
Visualization	2, 64	7.2	0.031	0.94	0.18	0.002**
Song	1, 32	4.55	0.090	1.00	0.13	0.041*
Visualization $\times$ song	2, 64	3.41	0.015	0.81	0.10	0.050

Note. GG correction is appropriate only for within-subject tests with more than one degree of freedom in the numerator.

* $p < 0.05$.

** $p < 0.01$.

*** $p < 0.001$.

ϵ = GG epsilon, p_{GG} = GG-corrected p values.

ratings for music), resulting in an aggregated mean externalization rating per participant for each source type condition. The corresponding results plot in Fig. 5(b) indicates slightly higher overall externalization ratings for phantom and moving sources. The ANOVA confirmed an effect of source type (see Table 4), and subsequent post hoc tests revealed that phantom sources [$t(32) = 5.34$, $p < 0.001$, 0.57 ± 0.12 vs. 0.47 ± 0.17] and moving sources [$t(32) = 5.95$, $p < 0.001$, 0.57 ± 0.14 vs. 0.47 ± 0.17] led to significantly higher externalization ratings than single sources, with no other significant pairwise comparisons.

Regarding the *single sources*, the analysis of the influence of the source type suggests that the AR visualization led to higher externalization for single sources [Fig. 5(c)] and that, in particular, the noise bursts presented from the RL virtual loudspeaker were perceived as more externalized than those presented by the other loudspeakers. The ANOVA yielded a significant main effect of visualization and loudspeaker but no interaction effect (see Table 4), confirming the observations from the plot, and post hoc comparisons for the factor visualization further supported the improved externalization through AR visualization, revealing that the AR visualization led to higher externalization ratings than display visualization [$t(32) = 3.03$, $p = 0.015$, 0.52 ± 0.24 vs. 0.45 ± 0.26] and no visualization [$t(32) = 2.63$, $p = 0.026$, 0.52 ± 0.24 vs. 0.45 ± 0.24], whereas there was no significant difference between display visualization and no visualization [$t(32) = 0.05$, $p = 0.964$, 0.45 ± 0.26 vs. 0.45 ± 0.24].

Right-tailed t tests comparing the AR condition with the display and no visualization condition for each sin-

gle source condition revealed significant differences for five of the eight comparisons (AR-C vs. Display-C, AR-L vs. Display-L, AR-RL vs. None-RL, AR-FUL vs. Display-FUL, AR-FUL vs. None-FUL), with the largest effect sizes for the C and FUL loudspeakers. However, when applying the Hochberg correction [63] for multiple comparisons, only the AR-C vs. Display-C condition remained significant [$t(32) = 3.00$, $p = 0.0026$, 0.44 ± 0.21 vs. 0.34 ± 0.25] (which is smaller than the adjusted alpha level of 0.0063), so some of the results for single sources might rather be interpreted as strong trends.

Finally, post hoc tests for the factor loudspeaker further confirmed the observations from the plot, revealing that the RL loudspeaker provided significantly higher externalization scores than all other loudspeakers (RL vs. C [$t(32) = 6.00$, $p < 0.001$, 0.60 ± 0.24 vs. 0.38 ± 0.24], RL vs. L [$t(32) = 5.65$, $p < 0.001$, 0.60 ± 0.24 vs. 0.45 ± 0.21], RL vs. FUL [$t(32) = 3.99$, $p = 0.001$, 0.60 ± 0.24 vs. 0.45 ± 0.26]). Furthermore, the analysis revealed that stimuli presented from the C loudspeaker resulted in significantly lower externalization than those from the L loudspeaker [$t(32) = 2.96$, $p = 0.017$, 0.38 ± 0.24 vs. 0.45 ± 0.21] and the RL loudspeaker (see results above).

The results plot for *phantom sources* in Fig. 5(d) suggests no clear influence of visualization, but some effect of phantom source position. In line with this, the ANOVA for the phantom source subset yielded a main effect only for the position factor, but no other main or interaction effect (see Table 4). Most interestingly, the post hoc comparisons revealed that phantom sources with $\phi = 115^\circ$ consistently led to significantly higher externalization ratings than

those with $\phi = 45^\circ$ ($(115^\circ, 0^\circ)$ vs. $(45^\circ, 0^\circ)$ [$t(32) = 8.32$, $p < 0.001$, 0.68 ± 0.20 vs. 0.46 ± 0.19], $(115^\circ, 0^\circ)$ vs. $(45^\circ, 20^\circ)$ [$t(32) = 4.58$, $p < 0.001$, 0.68 ± 0.20 vs. 0.52 ± 0.21], $(115^\circ, 20^\circ)$ vs. $(45^\circ, 0^\circ)$ [$t(32) = 6.74$, $p < 0.001$, 0.63 ± 0.20 vs. 0.46 ± 0.19], $(115^\circ, 20^\circ)$ vs. $(45^\circ, 20^\circ)$ [$t(32) = 3.09$, $p = 0.012$, 0.63 ± 0.20 vs. 0.52 ± 0.21]).

A look at the results for the *moving sources* in Fig. 5(e) also suggests no effect of visualization but a clear effect of the trajectory, with the vertical front trajectory yielding lower externalization ratings overall. Consequently, the ANOVA yielded a strong significant main effect of trajectory and no other main or interaction effects (see Table 4). Not surprisingly, post hoc tests showed that the stimuli on the vertical trajectory resulted in significantly lower externalization than those on the horizontal left [$t(32) = 6.80$, $p < 0.001$, 0.41 ± 0.22 vs. 0.64 ± 0.2] and horizontal right [$t(32) = 8.37$, $p < 0.001$, 0.41 ± 0.22 vs. 0.66 ± 0.22] trajectories.

Finally, inspection of the results plot for the *music* excerpts in Fig. 5(f) indicates a slight influence of the song and an effect of visualization, with a tendency for the AR visualization to yield higher externalization ratings. The ANOVA confirmed these observations, showing a significant main effect of visualization and song, but no significant interaction effect (see Table 4). Similar to the single sources, post hoc comparisons for the visualization factor showed significantly higher externalization ratings for the AR visualization compared to the display [$t(32) = 3.26$, $p = 0.005$, 0.60 ± 0.26 vs. 0.51 ± 0.27] and no visualization [$t(32) = 3.56$, $p = 0.004$, 0.6 ± 0.26 vs. 0.5 ± 0.26] but no significant difference between display and no visualization [$t(32) = 0.636$, $p = 0.53$, 0.51 ± 0.27 vs. 0.50 ± 0.26].

Hochberg-corrected right-tailed t tests comparing the AR condition with the display and no visualization condition for both songs revealed significant differences for all four comparisons (all $p < 0.015$), emphasizing the clear effect of AR visualization. The significant effect of song was because Billie Eilish's song yielded slightly higher externalization ratings than Elton John's song [$t(32) = 2.13$, $p = 0.041$, 0.58 ± 0.27 vs. 0.49 ± 0.26].

4 DISCUSSION

Binaural rendering of virtual loudspeaker setups has become a popular method for listening to and working with immersive audio, with perceived externalization being a key factor in creating a realistic spatial impression. This study examined the impact of AR, 2D display, and no visualization of a loudspeaker setup on the externalization of a 3DoF+ binaural rendering of a multichannel loudspeaker setup for different source types.

The results provide evidence that AR visualization of loudspeakers can significantly enhance the perceived externalization of the binaural reproduction compared to 2D visualization of the loudspeakers or no visualization. Notably, the main effect of visualization had a partial eta squared of 0.18, which is considered a medium to large effect in behavioral sciences [68], indicating that the impact of AR

visualization on externalization is substantial rather than subtle. However, the effect does not equally apply for all source types. It is more pronounced for signals reproduced by single virtual loudspeakers and for the immersive music excerpts reproduced by the entire virtual loudspeaker setup. It is less pronounced for phantom sources appearing between loudspeakers and for moving sources, where there was no significant effect for visualization, even though a trend towards higher perceived externalization with AR visualization can be observed for some source type variants.

Thus, enhancing externalization through AR visualization seems to work best when the acoustically perceived source direction matches the visual source location, so that the visual object acts as an anchor that pulls the percept farther away, similar to what is known from visual capture in localization [69]. In contrast, the lack of such an anchor for phantom and moving sources, where there is a mismatch between the visual and auditory location of the sound source, may be the reason why perceived externalization did not increase in these cases.

The significantly enhanced externalization of immersive music playback through AR visualization of the loudspeaker setup is particularly relevant for headphone-based immersive audio rendering and production tools. The effect, however, cannot be explained by a direct audiovisual consistency for individual sound sources, as observed for the noise bursts reproduced by single virtual loudspeakers, since typically, many sources in a multichannel mix are placed between loudspeakers using panning. Nevertheless, a higher audiovisual consistency is also given here when listeners see a virtual multichannel loudspeaker system for a spatial music mix, as this aligns with what the audio signal would naturally lead them to expect, provided they have gradually become accustomed and thus been sufficiently socialized to spatial music mixes played over loudspeakers.

A second effect evident in this study is the higher externalization for lateral sound sources than for sound sources in the median plane, with externalization increasing with lateralization. This applies to the signals presented by single virtual loudspeakers as well as for phantom and moving sources. An increase in perceived externalization with source lateralization was consistently found in previous studies [36, 38, 33, 35] and can most likely be explained by a lack of interaural differences for median plane sources. The absence of a significant interaction effect between source type variants and the visual conditions in this data (see the interaction terms in Table 4) strongly suggests that this is a purely acoustic effect. In line with this, when considering the single sources as an example, the rear left (RL) loudspeaker, which had the largest lateral angle, was perceived as more externalized, even though it was not in the participants' field of vision, and audiovisual consistency thus provides no explanation.

Furthermore, the aggregated results shown in Fig. 5(b) might give the impression that phantom sources were generally perceived as more externalized than single sources. However, this effect can be attributed to the variants selected within each source type. First, among the single sources, the

inclusion of the center loudspeaker significantly reduced the average externalization across the group. Second, the lateral phantom sources were positioned closer to the side of the head than the lateral single sources (45° vs. 30° and 115° vs. 135°), which tends to enhance the perceived externalization [36]. Both effects can be clearly seen in Figs. 5(c) and 5(d).

The different levels of externalization for the two immersive music mixes might also be due to an acoustic effect. In this experiment, a song by Billie Eilish (“I Didn’t Change My Number”) was rated as significantly more externalized than a song by Elton John (“Rocket Man”), which may be related to different mixing decisions. While in “Rocket Man” the center loudspeaker is the predominant source for the main vocals, the center speaker is used very little in the Billie Eilish mix. Since the main vocal is most often the most salient element of a pop song, it is possible that participants focused most on it, resulting in an overall lower externalization due to the predominance of the center speaker in the mix.

5 CONCLUSION

This study examined whether loudspeaker visualization enhances the perceived externalization of a binaurally rendered multiloudspeaker setup. In a listening experiment, participants listened to a 3DoF+ binaural rendering of a 7.0.4 loudspeaker setup and rated the perceived externalization under three visual conditions: (1) simulated loudspeakers as AR objects in the participant’s surroundings, (2) the simulated loudspeaker configuration on a 2D computer screen, and (3) no additional visual information. For each visual condition, participants were presented with noise bursts rendered as four different single loudspeaker sources, four different panned phantom sources, and three moving sources with different trajectories, as well as two immersive music mixes.

These findings demonstrate that AR visualization of virtual loudspeaker setups significantly enhances the perceived externalization of binaural rendering over headphones, both for single virtual loudspeaker sources and for immersive multichannel music mixes. This enhancement can be attributed to an increased audiovisual consistency. For a single virtual loudspeaker, this means that when the visualized loudspeaker location aligns with the acoustically perceived source direction, the visual object acts as a spatial anchor that pulls auditory perception away from the head. For immersive multichannel music mixes, audiovisual consistency might be achieved through AR visualization because seeing the acoustically simulated loudspeakers aligns with listeners’ expectations of spatial music mixes played over loudspeakers. Conversely, no significant improvement in externalization was observed for phantom sources between loudspeakers or for moving sources, where visual and auditory localization cues mismatch.

The results highlight the importance of multisensory integration, particularly in portable headphone-based immersive audio tools, where AR visualizations can ef-

fectively serve as spatial anchors to enhance the realism and externalization of binaural loudspeaker simulations. Future work on multisensory factors influencing perceived externalization should investigate whether the present findings, especially for immersive music, can be confirmed for room divergent conditions (i.e., studies are needed to investigate whether AR visualization of loudspeaker setups can also enhance externalization in conditions where the listener is in a different room than the one being acoustically simulated; a common use case with portable virtual studio software solutions). The results of such studies could provide further conclusive evidence for the effects found in this study and provide new insights into the fundamental multisensory processes involved in how object visualization affects perceived sound source externalization.

6 REFERENCES

- [1] Dolby Laboratories, “Getting Started With Dolby Atmos Music,” https://professionalsupport.dolby.com/s/article/Getting-Started-with-Dolby-Atmos-Music?language=en_US (accessed 22 Apr. 2025).
- [2] Fraunhofer IIS, “MPEG-H Audio,” <https://mpeg.com/> (accessed 22 Apr. 2025).
- [3] ITU-R, “Advanced Sound System for Programme Production,” *ITU-R Recommendation BS.2051-3* (2022 May). <https://www.itu.int/rec/R-REC-BS.2051-3-202205-I/en>.
- [4] ITU-R, “Multichannel Sound Technology in Home and Broadcasting Applications,” *ITU-R Recommendation BS.2159-9* (2022 Mar.). <https://www.itu.int/pub/R-REP-BS.2159-9-2022>.
- [5] Dear Reality, “dearVR MONITOR,” <https://www.dear-reality.com/products/dearvr-monitor> (accessed 22 Apr. 2025).
- [6] APL, “VIRTUOSO v2,” <https://apl-hud.com/product/virtuoso/> (accessed Apr. 22, 2025).
- [7] Noisemakers, “Binauralizer Studio 2,” <https://www.noisemakers.fr/product/binauralizer-studio-2/> (accessed 22 Apr. 2025).
- [8] P. Stadel, B. Bernschütz, and M. Rühl, “A Spatial Audio Impulse Response Compilation Captured at the WDR Broadcast Studios,” in *Proc. of the 27th Tonmeistertagung - VDT Intl. Conv.*, pp. 1–17 (Cologne, Germany) (2012 Nov.).
- [9] F. Brinkmann, A. Lindau, and S. Weinzierl, “On the Authenticity of Individual Dynamic Binaural Synthesis,” *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 1784–1795 (2017 Oct.). <https://doi.org/10.1121/1.5005606>.
- [10] A. Neidhardt and A. M. Zerlik, “The Availability of a Hidden Real Reference Affects the Plausibility of Position-Dynamic Auditory AR,” *Front. Virtual Real.*, vol. 2, paper 678875 (2021 Sep.). <https://doi.org/10.3389/frvir.2021.678875>.
- [11] B. Bernschütz, A. V. Giner, C. Pörschmann, and J. M. Arend, “Binaural Reproduction of Plane Waves With Reduced Modal Order,” *Acta Acust. united Acust.*, vol. 100, no. 5, pp. 972–983 (2014 Sep.). <https://doi.org/10.3813/AAA.918777>.

- [12] T. Lübeck, H. Helmholtz, J. M. Arend, C. Pörschmann, and J. Ahrens, "Perceptual Evaluation of Mitigation Approaches of Impairments Due to Spatial Under-sampling in Binaural Rendering of Spherical Microphone Array Data," *J. Audio Eng. Soc.*, vol. 68, no. 6, pp. 428–440 (2020 Jun.). <https://doi.org/10.17743/jaes.2020.0038>.
- [13] S. V. Amengual Garí, J. M. Arend, P. Calamia, and P. W. Robinson, "Optimizations of the Spatial Decomposition Method for Binaural Reproduction," *J. Audio Eng. Soc.*, vol. 68, no. 12, pp. 959–976 (2020 Dec.). <https://doi.org/10.17743/jaes.2020.0063>.
- [14] J. M. Arend, S. V. Amengual Garí, C. Schissler, F. Klein, and P. W. Robinson, "Six-Degrees-of-Freedom Parametric Spatial Audio Based on One Monaural Room Impulse Response," *J. Audio Eng. Soc.*, vol. 69, no. 7/8, pp. 557–575 (2021 Jul.). <https://doi.org/10.17743/jaes.2021.0009>.
- [15] P. Stade and J. M. Arend, "Perceptual Evaluation of Synthetic Late Binaural Reverberation Based on a Parametric Model," in *Proc. of the AES Intl. Conv. on Headphone Technology* (2016 Aug.), paper 3-3.
- [16] P. Stade, J. M. Arend, and C. Pörschmann, "Perceptual Evaluation of Synthetic Early Binaural Room Impulse Responses Based on a Parametric Model," presented at the *142nd Conv. of the Audio Engineering Society* (2017 May), paper 9688.
- [17] P. Coleman, A. Franck, P. J. Jackson, L. Remaggi, and F. Melchior, "Object-Based Reverberation for Spatial Audio," *J. Audio Eng. Soc.*, vol. 65, no. 1/2, pp. 66–77 (2017 Jan.). <https://doi.org/10.17743/jaes.2016.0059>.
- [18] D. Schröder and M. Vorländer, "RAVEN: A Real-Time Framework for the Auralization of Interactive Virtual Environments," in *Proc. of Forum Acusticum*, pp. 1541–1546 (Aalborg, Denmark) (2011 Jun.).
- [19] T. Wendt, S. van der Par, and S. D. Ewert, "A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation," *J. Audio Eng. Soc.*, vol. 62, no. 11, pp. 748–766 (2014 Nov.). <https://doi.org/10.17743/jaes.2014.0042>.
- [20] A. Lindau and S. Weinzierl, "On the Spatial Resolution of Virtual Acoustic Environments for Head Movements in Horizontal, Vertical and Lateral Direction," in *Proc. of the EAA Symposium on Auralization*, pp. 1–6 (Espoo, Finland) (2009 Jun.). <http://dx.doi.org/10.14279/depositonce-8668>.
- [21] J. De Muynke, D. Poirier-Quinot, and B. F. G. Katz, "Effect of Interpolation Artifacts on Perceived Stability of Nearby Sources in a Navigable Reverberant Virtual Environment," *J. Audio Eng. Soc.*, vol. 72, no. 10, pp. 664–678 (2024 Sep.). <https://doi.org/10.17743/jaes.2022.0158>.
- [22] A. Neidhardt and B. Reif, "Minimum BRIR Grid Resolution for Interactive Position Changes in Dynamic Binaural Synthesis," presented at the *148th Conv. of the Audio Engineering Society* (2020 May), paper 10371.
- [23] S. Werner, A. Neidhardt, F. Klein, and K. Brandenburg, "Comparison of Different Methods to Create an Interactive Augmented Auditory Reality Scenario Using Sparse Binaural Room Impulse Response Measurements," in *Proc. of the 44th DAGA*, pp. 1–4 (Munich, Germany) (2018 Mar.).
- [24] D. Bau, H. Himmelein, and C. Pörschmann, "Comparison of Non-Parametric Interpolation Techniques for Sparsely Measured Binaural Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 72, no. 7/8, pp. 479–492 (2024 Jul.). <https://doi.org/10.17743/jaes.2022.0150>.
- [25] S. Riedel and M. Frank, "Spatial Perception of Multi-Source Scenarios in Real and Virtual Loudspeaker Arrangements," in *Proc. of the 50th DAGA*, pp. 1–4 (Hannover, Germany) (2024 Mar.).
- [26] T. McKenzie, N. Meyer-Kahlen, R. Daugintis, et al., "Perceptually Informed Interpolation and Rendering of Spatial Room Impulse Responses for Room Transitions," in *Proc. of the 24th Intl. Congress on Acoustics (ICA)*, pp. 1–11 (Gyeongju, South Korea) (2022 Oct.).
- [27] L. McCormack, A. Politis, T. McKenzie, C. Hold, and V. Pulkki, "Object-Based Six-Degrees-of-Freedom Rendering of Sound Scenes Captured With Multiple Ambisonic Receivers," *J. Audio Eng. Soc.*, vol. 70, no. 5, pp. 355–372 (2022 May). <https://doi.org/10.17743/jaes.2022.0010>.
- [28] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording* (Springer, Cham, Switzerland, 2019). <https://doi.org/10.1007/978-3-030-17207-7>.
- [29] V. Pulkki, S. Delikaris-Manias, and A. Politis, *Parametric Time-Frequency Domain Spatial Audio* (Wiley, Hoboken, NJ, 2018). <https://doi.org/10.1002/9781119252634>.
- [30] D. Ackermann, J. Domann, F. Brinkman, et al., "Recordings of a Loudspeaker Orchestra With Multichannel Microphone Arrays for the Evaluation of Spatial Audio Methods," *J. Audio Eng. Soc.*, vol. 71, no. 1/2, pp. 62–73 (2023 Jan.). <https://doi.org/10.17743/jaes.2022.0059>.
- [31] M. Schutte, *Aspects of Room Acoustics, Vision and Motion in the Human Auditory Perception of Space*, Ph.D. thesis, Ludwig-Maximilians-Universität München, Munich, Germany (2021 Jul.). <https://doi.org/10.5282/edoc.28543>.
- [32] J. F. Sommerfeldt, J. M. Arend, F. Brinkmann, S. D. Ewert, and S. Weinzierl, "Source Directivity Implementation for Real-Time Room Acoustic Simulation With LiveRAZR," in *Proc. of the 50th DAGA*, pp. 1616–1618 (Hannover, Germany) (2024 Mar.).
- [33] V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopčo, "Sound Externalization: A Review of Recent Research," *Trends Hear.*, vol. 24, paper 2331216520948390 (2020 Jan.). <https://doi.org/10.1177/2331216520948390>.
- [34] G. Plenge, "On the Differences Between Localization and Lateralization," *J. Acoust. Soc. Am.*, vol. 56, no. 3, pp. 944–951 (1974 Sep.). <https://doi.org/10.1121/1.1903353>.
- [35] D. R. Begault and E. M. Wenzel, "Headphone Localization of Speech," *Hum. Factors*, vol. 35, no. 2, pp. 361–376 (1993 Jun.). <https://doi.org/10.1177/001872089303500210>.
- [36] W. O. Brimijoin, A. W. Boyd, and M. A. Akeroyd, "The Contribution of Head Movement to the Externalization and Internalization of Sounds," *PLoS*

ONE, vol. 8, no. 12, paper e83068 (2013 Dec.). <https://doi.org/10.1371/journal.pone.0083068>.

[37] 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," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–916 (2001 Oct.).

[38] T. Leclère, M. Lavandier, and F. Perrin, "On the Externalization of Sound Sources With Headphones Without Reference to a Real Source," *J. Acoust. Soc. Am.*, vol. 146, no. 4, pp. 2309–2320 (2019 Oct.). <https://doi.org/10.1121/1.5128325>.

[39] S.-M. Kim and W. Choi, "On the Externalization of Virtual Sound Images in Headphone Reproduction: A Wiener Filter Approach," *J. Acoust. Soc. Am.*, vol. 117, no. 6, pp. 3657–3665 (2005 May). <https://doi.org/10.1121/1.1921548>.

[40] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A Summary on Acoustic Room Divergence and Its Effect on Externalization of Auditory Events," in Proc. of the 8th Intl. Conv. on Quality of Multimedia Experience (QoMEX), pp. 1–6 (Lisbon, Portugal) (2016 Jun.). <https://doi.org/10.1109/QoMEX.2016.7498973>.

[41] J. Catic, S. Santurette, and T. Dau, "The Role of Reverberation-Related Binaural Cues in the Externalization of Speech," *J. Acoust. Soc. Am.*, vol. 128, no. 2, pp. 1154–1167 (2015 Aug.). <https://doi.org/10.1121/1.4928132>.

[42] H. G. Hassager, F. Gran, and T. Dau, "The Role of Spectral Detail in the Binaural Transfer Function on Perceived Externalization in a Reverberant Environment," *J. Acoust. Soc. Am.*, vol. 139, no. 5, pp. 2992–3000 (2016 May). <https://doi.org/10.1121/1.4950847>.

[43] J. Catic, S. Santurette, J. M. Buchholz, F. Gran, and T. Dau, "The Effect of Interaural-Level-Difference Fluctuations on the Externalization of Sound," *J. Acoust. Soc. Am.*, vol. 134, no. 2, pp. 1232–1241 (2013 Aug.). <https://doi.org/10.1121/1.4812264>.

[44] S. Li, R. Schlieper, and J. Peissig, "The Effect of Variation of Reverberation Parameters in Contralateral Versus Ipsilateral Ear Signals on Perceived Externalization of a Lateral Sound Source in a Listening Room," *J. Acoust. Soc. Am.*, vol. 144, no. 2, pp. 966–980 (2018 Aug.). <https://doi.org/10.1121/1.5051632>.

[45] W. O. Brimijoin and M. A. Akeroyd, "The Role of Head Movements and Signal Spectrum in an Auditory Front/Back Illusion," *i-Percept.*, vol. 3, no. 3, pp. 179–182 (2012 Apr.). <https://doi.org/10.1068/i7173sas>.

[46] E. Hendrickx, P. Stitt, J.-C. Messonnier, et al., "Influence of Head Tracking on the Externalization of Speech Stimuli for Non-Individualized Binaural Synthesis," *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. 2011–2023 (2017 Mar.). <https://doi.org/10.1121/1.4978612>.

[47] E. Hendrickx, P. Stitt, J.-C. Messonnier, et al., "Improvement of Externalization by Listener and Source Movement Using a 'Binauralized' Microphone Array," *J. Audio Eng. Soc.*, vol. 65, no. 7/8, pp. 589–599 (2017 Jul.). <https://doi.org/10.17743/jaes.2017.0018>.

[48] J. Udesen, T. Piechowiak, and F. Gran, "The Effect of Vision on Psychoacoustic Testing With Headphone-Based Virtual Sound," *J. Audio Eng. Soc.*, vol. 63, no. 7/8, pp. 552–561 (2015 Jul.). <https://doi.org/10.17743/jaes.2015.0061>.

[49] F. Klein, S. Werner, and T. Mayenfels, "Influences of Training on Externalization of Binaural Synthesis in Situations of Room Divergence," *J. Audio Eng. Soc.*, vol. 65, no. 3, pp. 178–187 (2017 Mar.). <https://doi.org/10.17743/jaes.2016.0072>.

[50] S. V. Amengual Garí, H. G. Hassager, F. Klein, J. M. Arend, and P. W. Robinson, "Towards Determining Thresholds for the Room Divergence Effect: a Pilot Study on Perceived Externalization," in Proc. of the Intl. Conv. on Immersive and 3D Audio (I3DA), pp. 1–7 (Bologna, Italy) (2021 Sep.). <https://doi.org/10.1109/I3DA48870.2021.9610835>.

[51] Physikalisch-Technische Bundesanstalt, "Absorption Coefficient Database," <https://www.ptb.de/cms/ptb/fachabteilungen/abt1/fb-16/ag-163/absorption-coefficient-database.html> (accessed 17 Jul. 2025).

[52] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950 (1979 Apr.). <https://doi.org/10.1121/1.382599>.

[53] F. Brinkmann, V. Erbes, and S. Weinzierl, "Extending the Closed Form Image Source Model for Source Directivity," in Proc. of the 44th DAGA, pp. 1298–1301 (Munich, Germany) (2018 Mar.).

[54] F. Brinkmann, H. Gamper, N. Raghuvanshi, and I. Tashev, "Towards Encoding perceptually Salient Early Reflections for Parametric Spatial Audio Rendering," presented at the 148th Conv. of the Audio Engineering Society (2020 Jun.), paper 10380.

[55] T. Jüterbock, F. Brinkmann, H. Gamper, N. Raghuvanshi, and S. Weinzierl, "Spatio-Temporal Windowing for Encoding Perceptually Salient Early Reflections in Parametric Spatial Audio Rendering," *J. Audio Eng. Soc.*, vol. 71, no. 10, pp. 664–678 (2023 Oct.). <https://doi.org/10.17743/jaes.2022.0099>.

[56] F. Brinkmann and S. Weinzierl, "AKtools - An Open Software Toolbox for Signal Acquisition, Processing, and Inspection in Acoustics," presented at the 142nd Conv. of the Audio Engineering Society (2017 May), eBrief 309.

[57] C. Borß and R. Martin, "An Improved Parametric Model for Perception-Based Design of Virtual Acoustics," in Proc. of the 35th AES Intl. Conv. (2009 Feb.), paper 3.

[58] J. M. Arend, T. Lübeck, and C. Pörschmann, "A Reactive Virtual Acoustic Environment for Interactive Immersive Audio," in Proc. of the AES Intl. Conv. on Immersive and Interactive Audio (IIA) (2019 Mar.), paper 9.

[59] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the Minimum Audible Difference in Direct-to-Reverberant Energy Ratio," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 450–461 (2008 Jul.).

[60] H.-P. Seraphim, "Untersuchungen Über die Unterschiedsschwelle Exponentiellen Abklingens von Rauschbandimpulsen," *Acta Acust. united Acust.*, vol. 8, no. 4, pp. 280–284 (1958 Jan.).

[61] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466 (1997 Jun.).

[62] C. Scheer, F. Lau, C. Sander, S. Weinzierl, and J. M. Arend, "Supplementary Material for 'The Influence of Visual Cues on the Sound Externalization of Virtual Loudspeaker Setups,'" *Zenodo*. (2025, Oct.). <https://doi.org/10.5281/zenodo.16274887>.

[63] Y. Hochberg, "A Sharper Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, vol. 75, no. 4, pp. 800–802 (1988 Dec.). <https://doi.org/10.1093/biomet/75.4.800>.

[64] G. V. Glass, P. D. Peckham, and J. R. Sanders, "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance," *Rev. Educat. Res.*, vol. 42, no. 3, pp. 237–288 (1972 Sep.). <https://doi.org/10.3102/00346543042003237>.

[65] S. W. Greenhouse and S. Geisser, "On Methods in the Analysis of Profile Data," *Psychometrika*, vol. 24, no. 2, pp. 95–112 (1959 Jun.). <https://doi.org/10.1007/BF02289823>.

[66] D. Cousineau, M.-A. Goulet, and B. Harding, "Summary Plots With Adjusted Error Bars: The *Superb* Framework With an Implementation in R," *Adv. Methods Pract. Psychol. Sci.*, vol. 4, no. 3, paper 25152459211035109 (2021 Aug.). <https://doi.org/10.1177/25152459211035109>.

[67] G. R. Loftus and M. E. J. Masson, "Using Confidence Intervals in Within-Subject Designs," *Psychon. Bull. Rev.*, vol. 1, no. 4, pp. 476–490 (1994 Dec.). <https://doi.org/10.3758/BF03210951>.

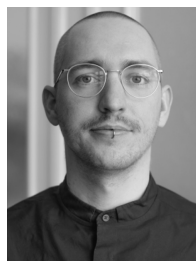
[68] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum Associates, Mahwah, NJ, 1988).

[69] D. Alais and D. Burr, "The Ventriloquist Effect Results From Near-Optimal Bimodal Integration," *Current Biol.*, vol. 14, no. 3, pp. 257–262 (2004 Feb.). <https://doi.org/10.1016/j.cub.2004.01.029>.

THE AUTHORS



Christian Scheer



Felix Lau



Christian Sander



Stefan Weinzierl



Johannes M. Arend

Christian Scheer is a research associate and Ph.D. candidate at the Audio Communication Group at TU Berlin. He received his B.Eng. degree in sound and video engineering from Hochschule Düsseldorf and Robert Schumann Hochschule Düsseldorf in 2021 and his M.Sc. degree in audio communication and technology from TU Berlin in 2024. His research interests include spatial audio, virtual acoustic realities, and psychoacoustics.

Felix Lau is a product owner at Dear Reality, where he oversees the development of various spatial audio software solutions. He holds a Master of Fine Arts as tonmeister for audiovisual media and began his career as a research and development engineer at the Institut für Rundfunktechnik, focusing on object-based audio. Today, Felix contributes not only to Dear Reality but also plays a role in other areas within the Sennheiser Group.

Christian Sander, cofounder and manager at Dear Reality (part of Sennheiser Group), is a renowned expert in the field of spatial audio technology and products. He received a Dipl.-Ing. degree in sound and video engineering from HS Düsseldorf and RSH Düsseldorf, Germany, in 2012 with a focus on binaural synthesis. He has worked as a sound engineer for music, broadcast, and postproduction, as well as an audio DSP developer.

Stefan Weinzierl is head of the Audio Communication Group at the Technische Universität Berlin. His research is focused on audio technology, virtual acoustics, room acoustics, and musical acoustics. With a diploma in physics and sound engineering, he received his Ph.D. in musical acoustics. He is coordinating a Master's program in audio communication and technology at TU Berlin and has coordinated international research consortia in the field of virtual acoustics (SEACEN) and music information retrieval (ABC.DJ).

Johannes M. Arend is an assistant professor at the Acoustics Lab, Department of Information and Communication Engineering, Aalto University, Finland. He leads the Technical Psychoacoustics research group, working at the intersection of technical developments in virtual acoustics and spatial audio, the perceptual evaluation of these technologies, and their application in hearing science studies on auditory and multisensory perception. Before joining Aalto University in 2024, he was a postdoctoral researcher at TU Berlin from 2022 to 2024 and a research associate at TH Köln from 2015 to 2022. He received a B.Eng. degree in media technology from HS Düsseldorf, Germany, in 2011; M.Sc. degree in media technology from TH Köln, Germany, in 2014; and Ph.D. degree (Dr. rer. nat.) from TU Berlin, Germany, in 2022.