DOI: 10.61782/fa.2023.0377

forum acusticum 2023

COMPARABLE SOUND SOURCE LOCALIZATION OF PLAUSIBLE AURALIZATIONS AND REAL SOUND SOURCES EVALUATED IN A NATURALISTIC EYE-TRACKING TASK IN VIRTUAL REALITY

S. Roßkopf[1] **L.O.H. Kroczek**[1] **F. Stärz**[2]
M. Blau[2, 4] **S. Van De Par**[3, 4] **A. Mühlberger**[1]
[1]Universität Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany
[2]Jade Hochschule Oldenburg, Ofener Str. 16, 26121 Oldenburg, Germany
[3] Carl von Ossietzky Universität Oldenburg, Carl-von-Ossietzky-Str. 9-11, 26129 Oldenburg, Germany
[4] Cluster of Excellence Hearing4All

ABSTRACT

Highly plausible audiovisual virtual scenes can be created using head-tracked binaural audio renderings presented via headphones combined with a visually-realistic scene presented via virtual reality (VR) glasses. Open questions are whether these plausible auralizations enhance social presence in VR and whether they allow sound source localization comparable to real sound sources. To address these questions, we implemented an eye-tracking paradigm in VR as naturalistic tool to measure spatial attention and sound source localization. In this study, 25 participants completed localization tasks and rated social presence and spatial audio quality. We compared three highly plausible auralizations to loudspeakers and to an anchor (gaming audio engine). Participants reported higher (almost 100%) externalization rates for all plausible auralizations compared to the anchor. Sound distance perception of plausible auralizations and loudspeakers do not differ. For azimuthal error, only for audio renderings based on individual HRTFs lower accuracy was found in comparison to the loudspeaker condition. Social presence was significantly higher in loudspeaker and plausible auralizations compared to the anchor condition. Furthermore, social presence and audio quality are strongly correlated. The implementation of audio renderings is therefore suggested for VR settings in which high levels of (social) presence are relevant (for example, VR exposure therapy).

*__Corresponding author__: sarah.rosskopf@ur.de

Keywords: *sound source localization, eye-tracking, virtual reality, presence, binaural headtracked auralizations*

1. INTRODUCTION

When creating a convincing virtual environment, the implementation of advanced auralizations is an obvious goal. There are various methods to synthesize acoustic virtual environments (in the following referred to as audio renderings), which all aim at creating a realistic spatial auditory impression [1]. In a recent study using a listening test, it was found that head-tracked binaural audio renderings were rated as close-to-real regarding acoustical properties such as reverberance, source distance, or overall quality [2]. In a current VR study, participants were not able to reliably distinguish between real loudspeakers placed in a real room and the corresponding head-tracked binaural audio renderings [3]. The authors used a seminar room scenario with a visually simulated room model presented via a head-mounted-display (HMD) and room-simulation-based audio renderings presented via headphones (vs. loudspeakers) and could confirm a convincing virtual seminar room scene, based on simulations. Creating a convincing spatial hearing impression via headphones is challenging, since not only source and listener dependent modifications, but also cognitive effects such as listeners' expectations e.g. derived from visual cues must be taken into consideration [1]. The question arises whether these plausible auralizations also allow close-to-real sound source localization in audiovisual VR. Furthermore, if a convincing virtual environment, e.g. a naturalistic seminar room scenario can be affirmed based on high plausibility, close-to-real auditory perceptions and sound source localization, what are the effects on presence? Presence is defined as subjective experience of "being there" [8], in

terms of feeling present in a virtual room. Another psychologically highly relevant variable of subjective experience in VR is *social presence*, which means the feeling of another person being there [8]. The extent to which plausible auralizations are needed to create social presence and perceived audio quality in VR is an open question. When it comes to creating a naturalistic task for sound source localization in a seminar room scenario, tracking of gaze behavior is an obvious approach. Humans tend to direct their gaze towards other people who are speaking and this behavior is modulated by audiovisual speech integration [4]. Speaker-directed gaze orientation is not only part of multimodal social attention but also seem to have perceptual advantages. Acoustic cues can be derived more accurately when presented in front to slightly lateral of the head [5]. Furthermore, directing gaze towards a sound enhances auditory spatial cue discrimination even when the head remains stationary [6]. Gaze behavior was also confirmed to be a useful measure of sound source localization [7]. With eye tracking paradigms, a naturalistic and implicit tool for measuring attentional resources is provided. In this VR study, we therefore used eye tracking to evaluate the impact of binaural audio renderings on sound source localization. We compared the plausible auralizations to real audio sources (loudspeakers) and to an anchor (VR engine implemented state-of-the-art 3D-audio-sound). The anchor was selected to have a base-line audio condition, that is commonly used in psychological VR research which also does incorporate room geometry, surface material, and head tracking. We therefore investigated whether accuracy of sound source localization of highly plausible binaural head-tracked audio renderings equals that of real sound sources and is superior to the state-of-the-art game-engine anchor. Furthermore, we were interested whether similar effects for subjective experience in terms of social presence and perceived spatial audio quality can be achieved and whether these dimensions are correlated. We hypothesized that sound source localization in all audio rendering conditions does not significantly differ from real loudspeakers and is more accurate than in the anchor condition. We furtherly hypothesized that for subjective experience (social presence and spatial audio quality) no differences between loudspeakers and all plausible auralizations can be found. We expect higher ratings of loudspeakers and plausible auralizations compared to the anchor and a correlation of social presence and spatial audio quality.

2. METHODS

2.1 Participants

Healthy adult individuals with self-reported unimpaired hearing, normal or corrected to normal vision, and German speaking experience of minimum 5 years were included in the study. Our sample ($N = 25$) consisted of 16 female and 9 male participants aged between 19 and 46 years ($M = 22.8$, $SD = 5.3$). All participants gave written informed consent. The study was in line with the Declaration of Helsinki and approved by the local ethics committee (University of Regensburg).

2.2 Room and visual virtual setup

The experiment took place in a seminar room of the University of Regensburg (room size: 10.6m x 7.1m x 3.3m). For the visual virtual room, we created a photorealistic model of the seminar room with the Unreal Game Engine (v 4.27, Epic Inc.) and Blender (v 2.79) using the exact room geometry and textures based on high-resolution photographs. The visual virtual environment was presented via a HMD (Vive Pro Eye, HTC). This device was also used for the measurement of eye-tracking data. For audiovisual virtual reality, an inaudible work station with passive cooling was used (Silentmaxx PC Kenko S-770i). The starting position of participants in the real room was matched to the according position in the visual virtual room model via an in-house-developed two-point calibration technique using custom-made mounts for the HTC motion controller. Data on our calibration technique were collected and a very high accordance of real and virtually visible positions could be affirmed.

Figure 1. Illustration of the visual virtual scene from participants' point of view.

Overall sixteen different female virtual agents were created using MakeHuman (v 1.2) and Blender (v 2.79). The virtual

agents were animated sitting on a chair with slight breathing movements. They were positioned to fill the whole auditorium, see fig 1. The position of eight agents exactly matched the (virtual) loudspeaker position in the room (four per experimental block, see below). The real and virtual loudspeakers were directed forward (parallel to the side walls) and accordingly all virtual agents directed their gaze straight forward. The position of agents' mouth was at 1.15m, which corresponded to the height of the acoustic center of loudspeakers. All agents wore a face mask, with the aim to reduce the interference of lacking lip movements on realism, as visual cues on speaker position were avoided. Participants were positioned in front of the auditorium at the lecturer position with a virtual visual notebook in front of them. At its display all instructions, rating scales and the vocabulary stimuli were presented.

2.3 Auditory setup

The virtual and real loudspeakers were positioned in front of the participants. In total, 8 different positions were used, four in each of the two experimental blocks. The height of the acoustical center of all loudspeakers was at 1.15m. Loudspeakers were placed at distances from 2.83m to 6.91m, and at azimuthal angles of 2° to 26°. The directivity pattern was taken into account and the loudspeaker was facing forwards (0°).

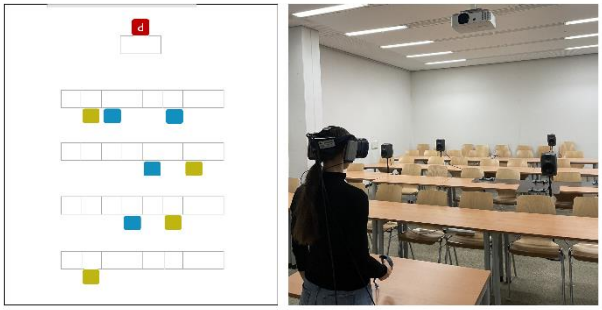

Figure 2. On the left, sound source positions are depicted. The blue squares illustrate positions of the first block, the green squares of the second block, the red square illustrates participants' start position. On the right, the setup in the seminar room is depicted.

We compared five different audio presentation modes. First, we used two-way active loudspeakers (Genelec 8030b, Genelec Oy, Isalmi, Finnland) as real sound sources in the room. All other audio conditions were presented using a headphone amplifier (Lake People G103P, Lake people electronic GmbH, Konstanz, Germany) and open headphones (AKG K1000, AKG Acoustics GmbH Vienna, Austria), which were adjusted to the HMD with custom-made 3D printed mountings [9]. Compared to circumaural headphones the spectral influence on the sound field produced by a real loudspeaker is smaller [9]. Nonetheless, an occlusion effect cannot be excluded for the use of K1000 headphones in comparison to real loudspeaker. An occlusion effect can also be assumed for the use of a HMD. However, no differences regarding plausibility could be found for audio renderings based on measurements with or without HMD [3]. For playback on loudspeaker and headphone, we used an external audio interface (RME Fireface, UC, Audio AG Haimhausen, Germany). Next, we used head-tracked binaural audio renderings based on three different BRIR sets, for which a high plausibility could be found [3]. The second audio condition, furtherly referred to as measHATS, were auralizations based on BRIR sets which were measured in the real room using a commercial head-and-torso-simulator (HATS; Kemar type 45BB, GRAS Sound and Vibration A/S, Holte, Denmark). The head-above-torso orientations of the HATS were varied between -90 and 90° in 5° steps, resulting in 37 azimuthal orientations. The elevation angle was fixed at 0°, the ear height was set to 1.60 m (lecturer position). This height was selected as approximation to the mean ear height of both, female and male participants. No adjustment of participants' ear height was made in this experiment, as the natural standing position in front of an auditorium was targeted. MEMS microphones (TDK type ICS-40619, TDK InvenSense, San Jose, CA, USA) inserted to the ear canals of the HATS using PIRATE ear plugs [10] were used for all measurements. BRIRs were measured using multiple exponential sweep stimuli (for further details see [2]). The third and fourth audio condition were auralizations based on BRIR sets which were simulated using RAZR (v 0.962b, [11]). The simulated room impulse responses were combined with measured head related impulse responses (HRIRs). In audio condition number three, furtherly referred to as simIndivHRIRs, individually measured HRIRs were used for the rendering of BRIRs. In audio condition number four, furtherly referred to as simHATS, generic HRIRs were used, measured with the above-described HATS. The measurement system for the HRIRs is a replication of the setup constructed and used at Jade Hochschule Oldenburg, for further details see [2]. Both simulated BRIRs were obtained for 37 azimuth angles (-90° to 90° in 5° steps) and nine elevation angles (-30° to 30° in 7.5° steps). Last, the fifth audio mode, furtherly called anchor, consisted of head-tracked binaural 3D auralizations created by a state of the art audio engine (Steam Audio v 4.1.4, Valve Corporation, Bellevue, WA, USA) implemented in the Unreal Engine. Real-time ray tracing

was used for modelling how sound was reflected by geometry, based on predefined acoustic material properties (e.g. tarpet for the floor). Occlusion and sound propagation was adapted towards the loudspeaker condition, in order to avoid salient loudness differences.
The stimuli, which were used equally often for each audio mode, consisted of 24 different dry recordings of female speech. Typical language course statements from one word (e.g. "station") to five word sentences (e.g. "What is it called in German?") were derived from a German learning program (studio21 A1 und A2, Cornelsen Verlag [12]). The stimuli were loudness normalized (integrated loudness function) in accordance with EBU R 128 and Hann windowed to overcome cutting artifacts. The order of stimulus presentation was pseudo-randomized via randomization lists. For presentation of stimuli, we created five different randomization lists, each beginning with a different audio mode (lists were counterbalanced across participants). All lists consisted of three blocks with 40 stimuli each. In each block, all variations of audio mode per sound position were included twice. The stimuli were pseudo-randomized within the blocks with following constraints: not more than three repetitions of same rendering, same position and same utterance.

2.4 Measurements and Data Processing

To measure sound source localization, participants were instructed to look at the location in the room where they assumed the sound source. Gaze behavior was recorded and analyzed during the task. For each trial, participants had to direct their gaze towards an object (which was not the laptop display or the room walls) within 3 s, otherwise the trial was repeated. If participants did not externalize a sound, which means that they perceived the sound inside their head, they were instructed to direct their gaze towards a blue button on the keyboard of the virtual notebook. Gaze behavior was analyzed offline using a custom Matlab script (v R2022a, The MathWorks, Inc., Natick, MA, USA) which categorized the gaze as fixation or saccade behavior. Fixations were defined using both velocity (< 75°/s) and gaze duration (> 200 ms) criteria [13]. Following a pre-registered analysis plan only the first fixation was analyzed. We used two measures for different aspects of sound source localization. First, we computed the angle (in deg °) between first fixation and sound source as indicator for azimuthal error. Second, we computed the deviance between the x-coordinate (longitudinal side) of the first fixation and sound source as indicator for the distance error. Azimuthal error and (absolute) distance error were averaged per audio condition and participant for statistical analyses. For the externalization index, the rate in % of externalized trials was computed per audio condition (and participant). Besides gaze behavior, the position of participants and sound sources were tracked. To measure subjective experience of virtual reality and audio scene, two 9-point Likert scaled ratings (1: "I disagree" – 9 "I agree.") were implemented within the scene. The first item concerned social presence (from German "Ich habe das Gefühl, dass gerade eine anwesende Person zu mir gesprochen hat." which translates as "I have the feeling that a person present has just spoken to me."). The second item concerned the perceived spatial audio quality (from German "Der Klang war so wie in einem Seminarraum." which translates as "The sound was like being in a seminar room."). For the tests of hypotheses, repeated measures ANOVAs were computed. We considered p-values < .05 as significant. If significant main effects of audio condition were found, post-hoc t-tests were computed. To prevent alpha-error-inflation, p-values corrected with the Bonferroni-Holm method are reported. For correlational analyses, Pearson's correlation coefficient is reported. For binomial data, Chi-Square Test of Independence are performed.

2.5 Procedure

The experiment consisted of two parts, which took place on different days. At the first appointment, informed written consent and questionnaires on demographic data were obtained, and individual HRIRs were measured. At the second appointment, the experiment was conducted. The presumed context for the presented audio-visual scene story was being in a language-learning course. After ensuring the approximately equal position of the HMD with respect to the Headphone equalization measurement (note that the headphones were fastened to the HMD), participants entered the seminar room "blindfolded" and were guided to the starting position by the experimenter and virtual footprints. This procedure allowed that participants remained unaware of the positions of loudspeakers. After calibration of eye-tracking, several practice trials were conducted to ensure understanding and manageability of tasks. Handling of ratings, the eye-tracking task and what to do when sound is internalized were practiced.

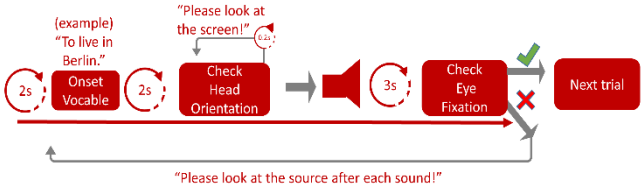

Figure 3. Illustration of the procedure of a trial.

The eye-tracking task was to look towards the spot where they assumed the sound source. When prepared, the first 60 trials of the experiments were run. These were followed by a break during which loudspeakers were rearranged. Possible auditory cues were masked by brown noise, which was played back on the headphones. Then, the next 60 trials were conducted. All trials started with the visual display of the vocabulary item (word or short sentence) on the notebook (see Figure 3). The orientation of the participant towards the notebook during the sound onset was controlled. If the rotation of the HMD exceeded 10°, a red text was displayed, instructing participants, to "Please look towards the screen." If verified, the sound was played back at the designated location. Head movements and gaze towards the source were encouraged as soon as the sound was played. The gaze behavior was recorded and analyzed for three seconds. If no valid pattern (no adjustment of gaze direction or fixation of the wall) was found, the trial was repeated. If the task was completed, the visual display of the vocabulary item disappeared and after an inter-trial interval of three seconds, the next trial started. After each sixth of the trials, the rating scales were presented in VR and had to be completed. After the VR experiment, participants were guided to the anteroom and again questionnaires on the experiment (difficulty of task, hypotheses, etc.) had to be answered.

3. RESULTS

3.1 Externalization

A Chi-Square Test of Independence was performed to assess the relationship between audio condition and rate of externalization. There was a significant relationship between audio condition and externalization, $\chi^2(4, 25) = 2204.00$, $p < .001$. Figure 4 illustrates, that this effect is driven by lower rates in the anchor condition.

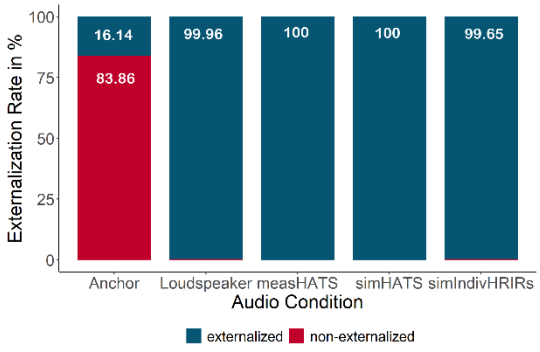

Figure 4. Rate of externalized trials in % as a function of audio condition.

3.2 Sound Source Localization

For trials in which a sound was not externally perceived, the sound source localization cannot be analyzed because participants directed their gaze towards a button on the notebook. The anchor condition had relatively low rates of externalization, resulting in only 68 analyzable trials over all participants ($M = 5.23$, $SD = 4.82$). For about half of participants ($N = 12$) no valid data for sound source localization of anchor stimuli was available. Therefore, all analyses concerning sound source localization were conducted with data only from the non-anchor audio conditions.

Azimuthal Error

A repeated measures ANOVA on the angle (in deg °) between first fixation and sound source as indicator for azimuthal error revealed a significant main effect of audio condition, $F(3,72) = 6.1$, $p = .009$, $\eta^2_G = 0.06$. Post-hoc paired t-tests revealed that in the simIndivHRIRs condition significantly higher azimuthal error could be found compared to the loudspeaker condition, $t(24) = -3.12$, $p = .028$, $d = -0.63$. Furtherly, in the simIndivHRIRs condition higher azimuthal error could be found compared to measHATS, $t(24) = -2.84$, $p = .045$, $d = -0.15$.

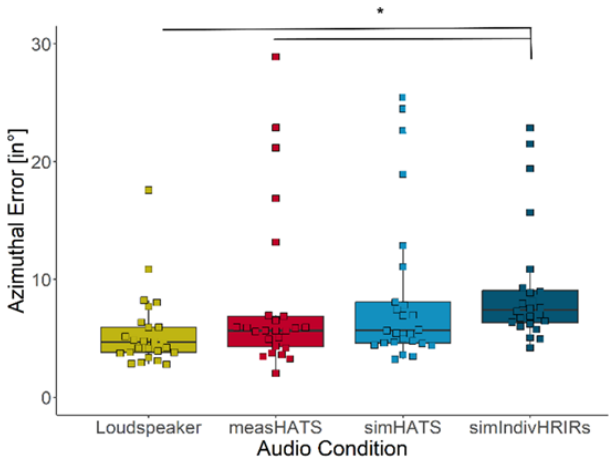

Figure 5. Deviance in ° between angle to sound source position and angle to fixated (estimated) position as indicator for azimuthal error as a function of audio condition.

Distance Error

A repeated measures ANOVA on the deviance between the x-coordinate (longitudinal side) of first fixation and sound source as indicator for distance error revealed no significant

main effect of audio condition, $F(3,72) = 1.2$, $p = .299$, $\eta_p^2 = 0.02$. Overall, participants tended to overestimate sound source distances.

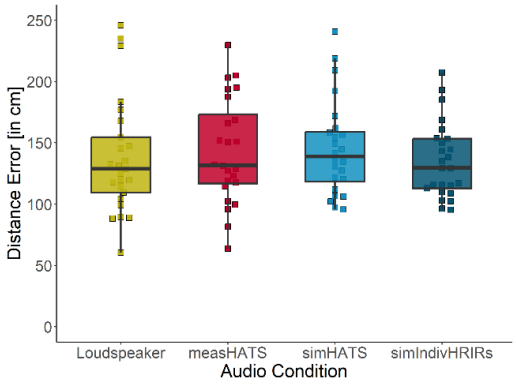

Figure 6. Deviance in cm between x-coordinate (longitudinal side) of first fixation and sound source as indicator for distance error as a function of audio condition.

3.3 Social Presence

A repeated measures ANOVA on the social presence ratings revealed a main effect of audio condition, $F(1.52, 36.41) = 46.2$, $p < .001$, $\eta_p^2 = 0.66$. Post-hoc paired t-tests revealed that for the anchor audio condition lower rating values than for all other audio conditions were found, all $t(24) < -6.5$, all $p < .001$, all $d < -1.3$. There were no significant differences between any of the other audio conditions.

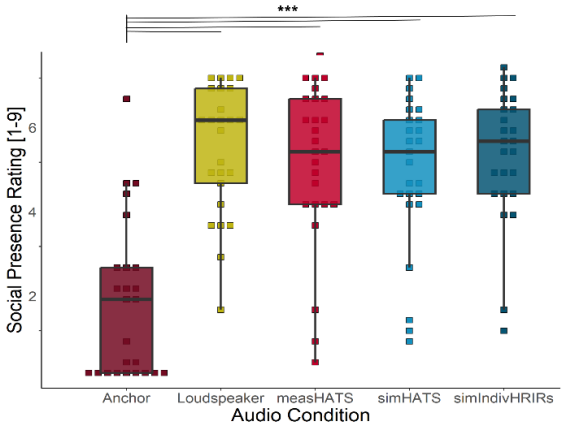

Figure 7 Social presence ratings ["I have the feeling that a person present has just spoken to me." 1 = "I disagree", 9 = "I agree"] as a function of audio condition.

3.4 Perceived spatial audio quality

A repeated measures ANOVA on the spatial audio quality ratings revealed a main effect of audio condition, $F(1.78, 42.70) = 32.2$, $p < .001$, $\eta_p^2 = 0.57$. Post-hoc paired t-tests revealed that for the anchor audio condition lower rating values than for all other audio conditions were found, all $t(24) < -5.5$, all $p < .001$, all $d < -1.1$. Furtherly, for stimuli based on simHATS spatial audio, quality was rated significantly lower than for simIndivHRIRs, $t(24) = -2.94$, $p = .042$, $d = -0.57$.

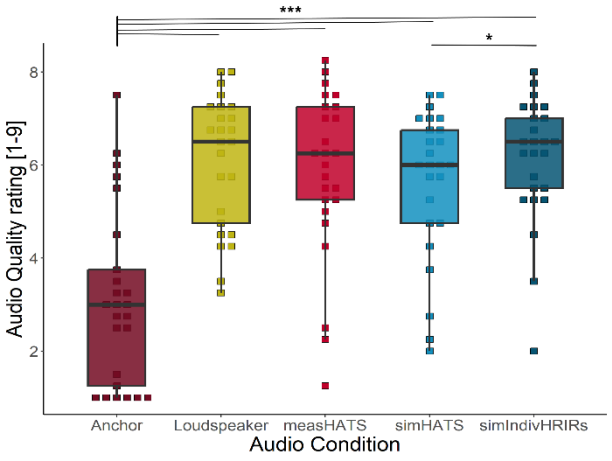

Figure 8. Spatial audio quality ratings ["The sound was like being in a seminar room." 1 = "I disagree", 9 = "I agree"] as a function of audio condition.

Correlation between social presence and spatial audio quality

The ratings of social presence and spatial audio quality were found to be strongly correlated $r(23) = .88$, $p < .001$. The relatively low ratings of the anchor condition are prominent. To rule out the possibility, that the correlation is only due to anchor vs. non-anchor conditions, correlational analyses without the anchor were conducted. Again, a strong correlation between social presence and spatial audio quality was found $r(23) = .84$, $p < .001$.

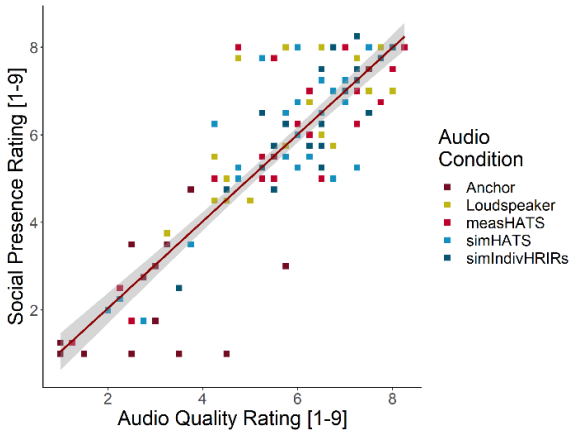

Figure 9. Spatial audio quality ratings ["The sound was like being in a seminar room." 1 = "I disagree", 9 = "I agree"] are strongly correlated with Social presence ratings ["I have the feeling that a person present has just spoken to me." 1 = "I disagree", 9 = "I agree"]. The colors indicate the audio condition.

4. DISCUSSION

Our results can be regarded as evidence for high quality of the here investigated plausible auralizations concerning externalization. In almost all trials (except two per loudspeaker condition and two per simIndivHRIRs) the sounds were externalized and could therefore be localized in the virtual room. In the anchor condition, only a rough fifth of trials were perceived outside the head. This is unexpected, since we used a state-of-the-art audio engine (Steam audio). One possible explanation for this finding is the lack of headphone equalization in the anchor condition. We used open headphones mounted to the HMD. This allowed direct and hidden comparison to the loudspeaker. For plausible auralizations we adjusted BRIRs in regard to the headphone-HMD-position in terms of individually measured and computed headphone equalizations. This could not be provided for the anchor, since the audio plugin is implemented in the gaming engine. An alternative explanation for the surprisingly poor results of the anchor could be the contrast to the simulations based on BRIRs which were precisely tailored towards the real room. In several test runs using only anchor stimuli, higher rates of externalization were found. Further experiments on context and contrast effects on externalization are being planned.

Overall, we were able to find comparable sound source localization for all (distance) or most (azimuthal) plausible auralizations and loudspeakers. For the perception of the distances of sound sources, no differences between real sound source and the plausible auralizations could be found. Concerning azimuthal detection of sound sources, for two audio rendering methods again no differences compared to the real sound source condition could be found. However, we found significant lower accuracy in the simIndivHRIRs condition. This is in contrast to typical findings, where individual HRTFs decreased azimuthal error when participants had to localize virtual sound sources [15]. One explanation could be that individual HRTF measurements are limited in precision. However, extensive efforts have been made to improve precision and reproducibility of measurements [3]. Further examinations of the individual HRTFs and possible explanations should be endeavored. Regarding the difference between simulations based on measured BRIRs and simulated BRIRs, a possible explanation is that in this experiment only in the simulated BRIR condition vertical head movements (elevation) were included. Furthermore, different ear heights of subjects were not taken into account. However, given the similarity of externalization and localization results between renderings and loudspeaker, it seems unlikely that this difference affected the present results. Comparable to previous studies, source distance was substantially overestimated [14]. Visual compression well known in the use of HMDs was proposed as underlying mechanism. Furtherly, means of azimuthal error are within the expected range [7, 14]. This can be seen as further evidence for the validity of the proposed eye-tracking paradigm used for measuring sound source localization.

Confirming our hypotheses, no differences concerning subjective experience ratings were found for loudspeaker vs. audio renderings. For social presence and perceived spatial audio quality significantly higher ratings were found for loudspeaker and audio rendering condition in comparison to the anchor. The levels of social presence are comparable to previous findings [14]. Interestingly, these two variables were strongly correlated. The correlation remained robust even when excluding the anchor condition (which prominently differed from the other conditions in terms of rating scores). This indicates that with a subjectively higher quality of the audio rendering in a VR scene higher levels of social presence can be induced (or vice versa). However, it is also possible that participants referred to a related construct, although the two rating items were formulated very differently. A comprehensive investigation of the interplay of audio quality and social presence is pending, here also inverted items should be implemented. All in all, the results can be seen as an impulse to implement sophisticated auralizations in the field

of clinical psychology, since higher levels of presence supporting smoother social interactions may contribute to enhanced emotional processing. This can advance research in socio-emotional settings or psychotherapeutic interventions using VR.

5. ACKNOWLEDGMENTS

This work is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under the project ID 422686707, SPP2236 – AUDICTIVE – Auditory Cognition in Interactive Virtual Environments [16]. We would like to specially thank Marieke Bruckmann and Nora Schmid for their help with data acquisition and to Andreas Ruider and Alexander May for their technical support.

6. REFERENCES

[1] K. Brandenburg, F. Klein, A. Neidhardt, and S. Werner. „Auditory Illusion over Headphones Revisited." *The Journal of the Acoustical Society of America*, vol. 141, no. 5, 2017.

[2] M. Blau, A. Budnik, M. Fallahi, H. Steffens, S.D. Ewert, and S. Van de Par. "Toward realistic binaural auralizations–perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario." *Acta Acustica* , vol. 5, no. 8, 2021.

[3] F. Stärz, L.O.H. Kroczek, S. Roßkopf, A. Mühlberger, S. Van de Par, and M. Blau. "Perceptual comparison between the real and the auralized room when being presented with congruent visual stimuli via a head-mounted display." In *Proc. Of the 24th International Congress on Acoustics*, Gyeongju, Korea, 2022.

[4] T. Foulsham, and L.A. Sanderson. "Look who's talking? Sound changes gaze behaviour in a dynamic social scene." *Visual Cognition*, vol. 21, no. 7: 922-944, 2013.

[5] J. C. Middlebrooks, and Z. A. Onsan, "Stream segregation with high spatial acuity." *J. Acoust. Soc. Am.*, vol. 132, no. 6, 3896-3911, 2012.

[6] R. K. Maddox, D. A. Pospisil, G. C. Stecker, and A. K. C. Lee, „Directing Eye Gaze Enhances Auditory Spatial Cue Discrimination." *Current Biology*, vol 24, 748-752, 2014

[7] R. Schleicher, S. Spors, D. Jahn, and R. Walter „Gaze as a measure of sound source localization." in *Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*, Pitea, Sweden, 2010.

[8] J. Diemer, G. W. Alpers, H. M. Peperkorn, Y. Shiban, and A. Mühlberger, "The impact of perception and presence on emotional reactions: a review of research in virtual reality." *Frontiers in psychology*, vol. 6, no. 26., 2015.

[9] F. Stärz, L. O. H. Kroczek, S. Roßkopf, A. Mühlberger, S. Van de Par, & M. Blau. (2023a). Mounting extra-aural headphones to a head-mounted display using a 3D-printed support. DAGA 2023, Hamburg, 1636–1639.

[10] F. Denk, F. Brinkmann, A. Stirnemann, and B. Kollmeier, „The PIRATE: an anthropometric earPlug with exchangeable microphones for Individual Reliable Acquisition of Transfer functions at the Ear canal entrance", in *Jahrestagung für Akustik (DAGA),* Rostock, Deutschland, 2019.

[11] T. Wendt, S. Van de Par, and S. D. Ewert. "A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation." *Journal of the Audio Engineering Society*, vol. 62, no. 11, 748–766, 2014.

[12] H. Funk, B. Lex, and B. Redecker, "Studio [21]: Deutsch als Fremdsprache. Das Deutschbuch.", *Cornelsen*, 2019.

[13] D. D. Salvucci, and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols." *In Proceedings of the 2000 symposium on Eye tracking research & applications*, 71-78, 2000.

[14] S. Roßkopf, L.O.H. Kroczek, F. Stärz, M. Blau, S. Van de Par, and A. Mühlberger. The Effect of Audio-Visual Room Divergence on the Localization of Real Sound Sources in Virtual Reality. DAGA 2023, Hamburg, 1431–1434.

[15] D. R. Begault, and E. M. Wenzel, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source.", *Journal of the Audio Engineering Society. Audio Engineering Society,* vol. 49, no. 10, 904-9416, 2001.

[16] DFG project homepage, URL: https://gepris.dfg.de/gepris/projekt/422686707