Influence of audio rendering in virtual environments on realism, presence, and socio-cognitive processing

Matthias Blau[1, 4], Steven van de Par[2, 4], Andreas Mühlberger[3], Felix Stärz[1], Leon O. H. Kroczek[3], Sarah Roßkopf [3]

[1] *Institut für Hörtechnik und Audiologie, Jade Hochschule Wilhelmshaven/Oldenburg/Elsfleth, D-26121 Oldenburg, Email {matthias.blau}{felix.staerz}@jade-hs.de,* [2] *Acoustics Group, Carl von Ossietzky Universität Oldenburg, D-26111 Oldenburg, Email steven.van.de.par@uni-ol.de,* [3] *Department of Psychology, Clinical Psychology and Psychotherapy, Universität Regensburg, D-93053 Regensburg, Email {andreas.muehlberger}{leon.kroczek}{sarah.rosskopf}@uni-regensburg.de,*
[4] *Cluster of Excellence Hearing4All*

Introduction

Although it is commonly agreed that auditory stimuli are crucial in studying perception and cognition in interactive virtual environments (iVEs), the quality of the audio rendering is often neglected.

In this project, we create and evaluate virtual representations of communication scenes in mid-sized rooms, by using a combination of visual renderings presented via head-mounted displays (HMDs) and head-tracked binaural rendering presented via headphones.

We first aim at determining the *requirements* regarding the (binaural) audio rendering *for a high degree of realism* in iVEs. Once it is known how to achieve highly realistic virtual scenes, more specific questions can be addressed, e.g.: Does the degree of realism *influence socio-cognitive processing such as social presence and the elicitation of social anxiety* in communication scenes? Is the achievable degree of realism sufficient to *reliably analyze the effect of room-acoustical modifications* in iVEs?

In the following, an overview of attempts towards answering these questions will be given.

From audio-only to audiovisual iVEs: binaural audio rendering for a high degree of realism

Initially, the question of how to arrive at a high degree of realism in iVEs using visual presentation via state-of-the-art HMDs and binaural audio rendering via headphones was addressed. Starting with a reference scene (speech stimuli presented via loudspeakers in the real room, see fig. 1), we first substituted the audio by a virtual presentation via headphones and then substituted the visual information by a virtual presentation via HMDs as well, thus effectively arriving at a virtual audiovisual scene.

In the first experiment (Blau et al., 2021), the reference scene was compared to the virtual (head-tracked) audio rendering via headphones, while the participant was present in the real room, see fig. 1. In other words, everything was real except the binaural audio rendering. Participants were instructed to put the headphones off and on (the audio was rendered to the loudspeaker when the headphones were taken off and to the headphones when they were on) and to rate the agreement between loudspeaker and headphone presentation with respect to different perceptual attributes.

Very good agreements between the loudspeaker presentation and the binaural rendering were found (see fig. 1), with only small differences between perceptual attributes and loudspeaker positions, under the following conditions:

- the directivity of the loudspeakers had to be taken into account, either directly by measuring BRIRs or indirectly by incorporating the loudspeaker directivity into the room simulation,
- for the room simulation, an improved version of the RAZR package (Wendt et al., 2014) was used, in which image sources up to third order and a feedback delay network were employed and absorption coefficients were adapted to match the measured monaural reverberation time,
- headphones were individually equalized.

It must be noted that the other premises (head-tracked binaural rendering, speech stimuli) add to this list of requirements.

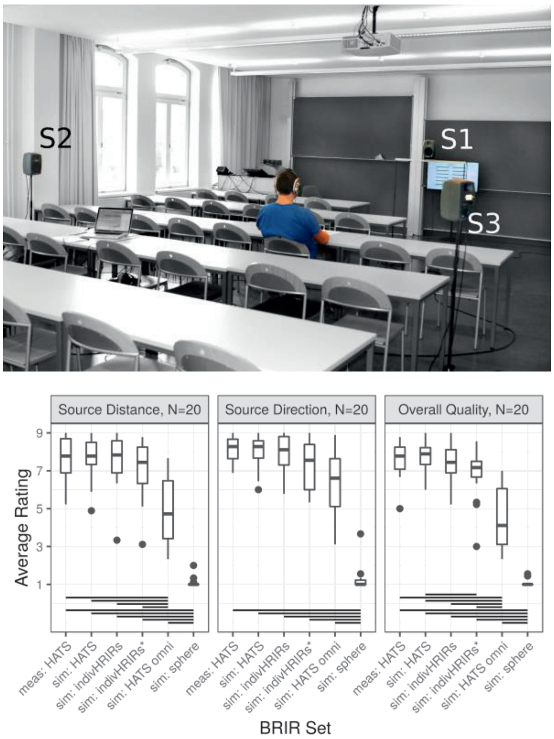

Figure 1: Experimental set-up and selected results from (Blau et al., 2021), figs. 1 and 6. **Top:** classroom scenario with three loudspeaker sources (S1..S3) and participant wearing headphones. **Bottom:** rated agreement between loudspeaker and headphone presentation for selected perceptual attributes.

Interestingly, both measured and simulated binaural room impulse responses (BRIRs) yielded highly realistic renderings, and individual head-related transfer functions (HRTFs) did not result in a benefit over generic (head-and-torso simulator, HATS) HRTFs.

A shortcoming of the first experiment was that the headphones had to be taken off to listen to the loudspeakers in the room. Hence, the reference situation was explicitly known to the participants, which could have biased their judgments. In order to confirm the requirements for highly realistic binaural audio renderings in a blind comparison to the reference situation, we developed a setup featuring headphones hovering over the ears (Stärz et al., 2023a), see also fig. 2. In addition, the visual impression was substituted by a virtual presentation of the classroom scene via a state-of-the-art HMD, while the participant wearing the HMD was still in the real room. The audio rendering could be delivered via the loudspeakers in the room or via the head-tracked binaural rendering described above. The participants simply had to decide whether the rendering they were listening to originated from a loudspeaker or not. Using signal detection theory, the fraction of correct answers was transformed into a sensitivity value d'. We have chosen d'=1 as threshold for reaching perceptual significance because it corresponds closely to the level of detectability that is obtained in typical 2-down 1-up adaptive staircase methods that converge at 71% correct (assuming e.g. a 1-IFC or ABX procedure). The results (fig. 2) indicated that the requirements listed above yield highly plausible renderings – only the rendering in which the true source directivities were substituted by omnidirectional ones yielded d' values greater than one (Stärz et al., 2022). In this experiment, the use of individual HRTFs resulted in a lower median sensitivity compared to the use of HATS HRTFs, but this difference was not statistically significant. It was also remarked that the visual model had to be carefully calibrated to avoid localization mismatch between the real loudspeakers and the audiovisual rendering.

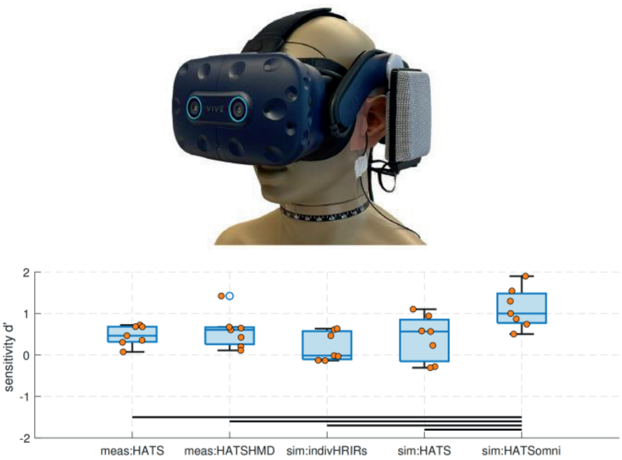

Figure 2: Setup and results from the blind comparison between loudspeaker and head-tracked binaural renderings. **Top:** combination of HMD and headphones hovering over the ears (Stärz et al., 2023a, fig. 2). **Bottom:** resulting sensitivity values for different binaural audio renderings (Stärz et al., 2022, fig. 3).

From visual-only to audiovisual iVEs: influence of audio rendering on distance perception and social presence

As a first approach to study auditory distance perception in Virtual Reality, an experiment was conducted to test the influence of affective voice on sound localization (Kroczek et al., 2022). Participants (N = 44) were placed in a virtual classroom visually presented via HMD, while speech stimuli (angry vs. neutral voices) were presented via loudspeakers (position of loudspeakers was not visible). A placement task was implemented where participants were asked to position a virtual agent on the perceived sound source using a motion controller (see fig 3). Interestingly, our results showed that angry voice stimuli were placed at greater distances than neutral voice stimuli, suggesting an influence of top-down expectations regarding affective dimensions of auditory stimuli for sound distance perception. Furthermore, the experiment allowed to introduce the placement task as a valid technique to study sound distance perception of social stimuli in VR.

Since in the experiment the visually displayed room differed from the room in which the experiment was performed, the extent to which the visual presentation of a room via HMD influences auditory spatial perception was investigated next (Roßkopf et al., 2022). Again real loudspeakers in the room were used as sound sources, the visual representation of the room was presented via HMD, and localization was measured via the placement task. We manipulated room (in)congruency and measured accuracy of sound source localization. Participants (N = 30) yielded lower accuracy of localization and higher rates of front back confusions when an incongruent visual room model was displayed via HMD. Interestingly, in both room conditions participants overestimated sound distances. This could be due to the visual compression resulting from the use of HMDs (Renner et al., 2013).

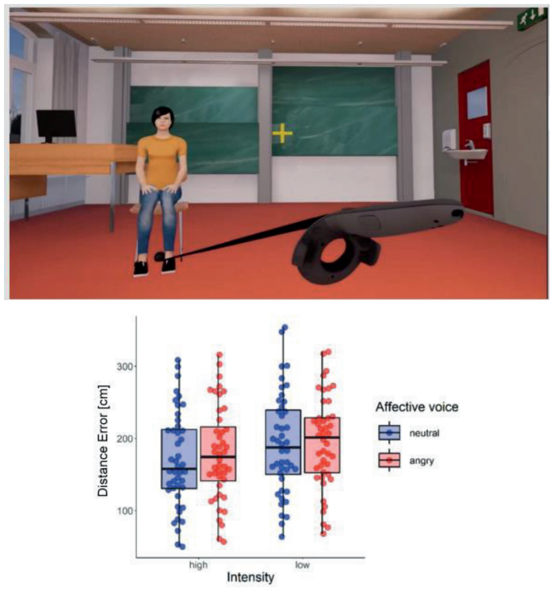

Figure 3: Placement task and selected results from Kroczek et al. (2022) **Top**: Placement of virtual agent using a motion controller in a virtual classroom scenario. **Bottom**: Effect of affective voice on sound localization.

Figure 4: Duplication of the HRTF measurement setup at the Department of Psychology, Clinical Psychology and Psychotherapy, University of Regensburg.

In order to confirm the previous findings on the similarity of head-tracked binaural renderings and loudspeaker with a broader variety of measurement techniques and variables concerning clinical psychology research, the HRTF measurement setup at Jade Hochschule Oldenburg was duplicated and installed at the Institute of Psychology at the University of Regensburg, see fig. 4. This allows the comparison of individually measured HRTFs with generic HRTFs in regard to different psychological and psychophysiological measurements. Two experiments were conducted investigating the effect of audio renderings on sound source localization, social presence, and subjective spatial audio quality in VR. Comparable to a previous study (Stärz et al., 2022) participants (N = 25) were present in a seminar room while its visually rendered model was displayed on a HMD on which open headphones were adjusted (fig. 2). This allowed the direct and hidden comparison of audio renderings to real loudspeakers in the room. In the first experiment, again the before-described placement task was used, whereas in the second experiment, a naturalistic eye-tracking task was used to measure sound source localization (Roßkopf et al., 2023b). We compared simulated BRIRs based on individual HRIRs, simulated BRIRs based on HATS, measured BRIRS based on HATS with loudspeakers in the room, and an anchor condition (state of the art audio engine implemented in the virtual reality gaming engine). Again, a high accordance of all binaural head-tracked audio renderings with the loudspeakers could be found regarding localization accuracy, but also regarding social presence. To add, there was no superiority of individualized HRIRs. Compared to the anchor, higher levels of social presence could be elicited in participants and social presence was strongly correlated to perceived spatial audio quality. This indicates, that with higher quality of audio renderings higher levels of social presence can be gained in a virtual seminar room scenario. This can be very useful when it comes to applications in the field of clinical psychology such as virtual reality exposure therapy of social anxiety.

Perception transfer from the real room to a different reproduction room

With the availability of highly realistic audiovisual iVEs, the question of how to transfer perceptual results obtained in a real room to a different reproduction room can be approached.

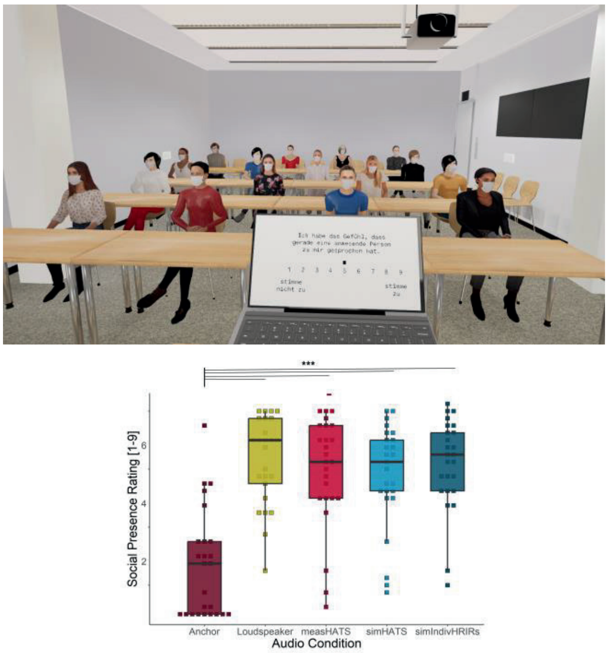

Figure 5: Experimental setup and selected results from eyetracking experiment (Roßkopf et al., 2023b). **Top**: Eyetracking Task for sound source localization and social presence rating. **Bottom**: Social Presence Ratings as a function of Audio Condition.

As an example application, one may think of rating room-acoustical attributes of different design variants of a performance space to be constructed via audiovisual VR in an office room or a hearing booth. Since, strictly speaking, no external reference exists in the reproduction room, experimental paradigms that rely on such a reference, such as ABX-tests for authenticity or yes-no-tasks for plausibility can't be used. Instead, rating scales for representative room-acoustical attributes were used in the following. Besides the obvious question which attributes are representative, it is important to know whether participants can reliably rate these attributes and to which degree the ratings are independent of the reproduction room.

In Stärz et al. (2023b), participants had to rate five room-acoustical attributes while listening to speech stimuli in the seminar room shown in fig. 1, wearing the HMD+headphone setup shown in fig. 2. Four of the attributes were taken from RAQI (Weinzierl et al., 2018). In addition, the overall reproduction quality was rated, which was defined as the room-acoustical quality of the room the participants were listening to. There were six different variants of the audio rendering: one rendering via the real loudspeaker in the room and five head-tracked binaural renderings. In two of the latter, the room simulation was altered by increasing and decreasing all wall absorption coefficients by 15 %, respectively. This resulted in similar changes in the measured reverberation times and of about 0.3 dB in the A-weighted level. A multiple-stimulus rating paradigm with the possibility to re-order the items according to the current ratings was used. Selected results are shown in fig. 6: the overall reproduction quality was rated high for three of the simulation-based binaural renderings, but not as high for the real loudspeaker, the simu-

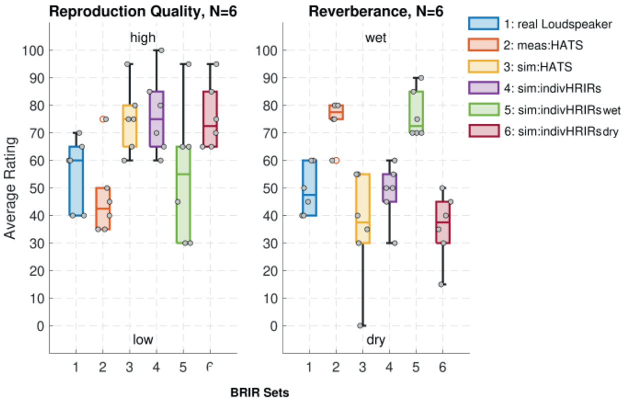

Figure 3: Rating of room-acoustical attributes in a highly realistic audiovisual VE. Modified from (Stärz et al., 2023b), fig. 3.

lation based on measured BRIRs and the simulation with increased reverberation. These ratings can to a large degree be explained by reverberance, except for the real loudspeaker. The increased reverberation was rated as markedly "wetter" and also as louder (not shown in fig. 6) than the unaltered variant, which indicates that subtle differences between variants can reliably be spotted by expert listeners in a multiple stimulus paradigm. It remains to be seen whether these results are confirmed when the experiment is carried out in a different reproduction room.

Conclusions and outlook

For speech stimuli, highly realistic audiovisual renderings can be achieved by using visual presentations via state-of-the-art HMDs and head-tracked binaural renderings of computer-generated scenes. For the highest degree of realism, the room simulation (RAZR: image sources up to 3rd order, plus feedback delay network) should be adapted to match the monaurally measured reverberation time of the room, source directivity should be taken into account, and individual headphone equalization should be employed. The use of individual HRTFs, on the other hand, appears to add only a small advantage, if any.

Using the high-realism audio rendering together with visual presentation via state-of-the-art HMDs, expert listeners located in the real room are able to discern subtle differences (in particular, in reverberance and loudness) between virtual room variants in a multiple-stimulus rating paradigm.

HMDs introduce a visual distance compression which may lead to a discrepancy between the perceived visual and auditory distance in audiovisual VR, depending on the experimental paradigm. In addition, top-down expectations regarding affective dimensions of auditory stimuli and audiovisual (in)congruency influence distance perception. There is a strong correlation between the quality of the binaural audio rendering and perceived social presence in a virtual seminar room scenario which may have implications for socio-cognitive research as well as for applications like psychotherapy, e.g. in virtual reality exposure therapy of social anxiety.

Further work is envisaged on the influence of the reproduction room on room-acoustical ratings, on the extension of highly realistic binaural renderings to other relevant stimuli such as social noise or music, on the handling of virtual agents as social interactions partners, and on efficient attention switching in virtual social situations with multiple interaction partners.

Acknowledgement

This work was funded by Deutsche Forschungsgemeinschaft (DFG, German research Foundation), project ID 444832396, within SPP 2236 Audictive.

References

Blau, M., Budnik, A., Fallahi, M., Steffens, H., Ewert, S. D., & van de Par, S. (2021). Toward realistic binaural auralizations—Perceptual comparison between the auralized and the real room for a classroom scenario. Acta Acustica, 5, 8.

Kroczek, L. O. H., Rosskopf, S., Stärz, F., Blau, M., van de Par, S., & Mühlberger, A. (2022). The influence of affective voice on sound distance perception. DAGA 2022, Stuttgart, 1163–1166.

Renner, R. S., Velichkovsky, B. M., & Helmert, J. R. (2013). The perception of egocentric distances in virtual environments - a review. ACM Computing Surveys (CSUR), 46(2), 1-40.

Roßkopf, S., Kroczek, L. O. H., Stärz, F., Blau, M., van de Par, S., & Mühlberger, A. (2023a). The Effect of Audio-Visual Room Divergence on the Localization of Real Sound Sources in Virtual Reality. DAGA 2023, Hamburg, 1431–1434.

Roßkopf, S., Kroczek, L. O. H., Stärz, F., Blau, M., van de Par, S., & Mühlberger, A. (2023b). A Naturalistic Eye-Tracking Approach to Evaluate the Influence of Auralizations on Sound Source Localization and Social Presence in Virtual Reality. Proc. Forum Acusticum 2023, Torino (Italy).

Stärz, F., Kroczek, L. O. H., Rosskopf, S., Mühlberger, A., van de Par, S., & Blau, M. (2022). Perceptual comparison between the real and the auralized room when being presented with congruent visual stimuli via a head-mounted display. Proc. 24th ICA, Gyeongju (Korea).

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

Stärz, F., Kroczek, L. O. H., Roßkopf, S., Mühlberger, A., van de Par, S., & Blau, M. (2023b). Comparing room acoustical ratings in an interactive virtual environment. Proc. Forum Acusticum 2023, Torino (Italy).

Weinzierl, S., Lepa, S., & Ackermann, D. (2018). A measuring instrument for the auditory perception of rooms: The Room Acoustical Quality Inventory (RAQI). JASA, 144(3), 1245–1257.

Wendt, T., van de Par, S., & Ewert, S. D. (2014). A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation. JAES, 62(11), 748–766.