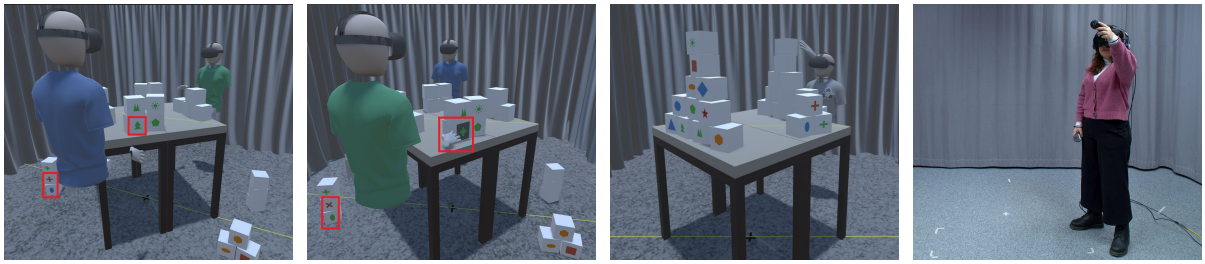
Evaluating the Effect of Binaural Auralization on Audiovisual Plausibility and Communication Behavior in Virtual Reality

Felix Immohr *
Audiovisual Technology Group, Technische Universität Ilmenau, Germany
Gareth Rendle †
Virtual Reality and Visualization, Bauhaus-Universität Weimar, Germany
Anton Lammert ‡
Virtual Reality and Visualization, Bauhaus-Universität Weimar, Germany
Annika Neidhardt §
Institute of Sound Recording, University of Surrey, UK
Victoria Meyer zur Heyde ¶
Electronic Media Technology Group, Technische Universität Ilmenau, Germany
Bernd Froehlich ‖
Virtual Reality and Visualization, Bauhaus-Universität Weimar, Germany
Alexander Raake **
Audiovisual Technology Group, Technische Universität Ilmenau, Germany



(a) (b) (c) (d)

Figure 1: In a conversation test investigating aspects of audiovisual plausibility and communication behavior, participants collaborated to find differences between sets of shapes, as shown in (a) and (b). The task was performed under two auralization conditions and two scene arrangement conditions, distributed (DIST, shown in (a) and (b)) and shadowing (SHAD, shown in (c)). The virtual room was a replica of the real room (d).

Abstract

Spatial audio representations have been shown to positively impact user experience in traditional, non-immersive communication media. While spatial audio also contributes to presence in single-user immersive VR, its impact in virtual communication scenarios has not yet been fully understood. This work aims to further investigate which communication scenarios benefit from spatial audio representations. We present a study in which pairs of interlocutors undertake a collaborative task in an audiovisual Virtual Environment (VE) under different auralization and scene arrangement conditions. The novel task is designed to encourage simultaneous conversation and movement, with the aim of increasing the relevance of spatial hearing. Results are obtained through questionnaires measuring social presence and plausibility, as well as through conversational and behavioral analysis. Although participants are shown to favor binaural auralization over diotic audio in a direct active-listening comparison, no significant differences in social presence, plausibility, or communication behavior could be found. Our results suggest that spatial audio may not affect user experience in dyadic communication scenarios where spatial auditory information is not directly relevant to the considered task.

Index Terms: Human-centered computing—Human computer interaction (HCI)—HCI and evaluation methods—User studies; Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality;

1 Introduction

Spatial audio may be realized using loudspeakers or through binaural rendering techniques with headphones that leverage interaural differences to create impressions of localizable sound sources [31], contrasting with diotic audio, where no interaural differences exist. Positive effects of spatial audio have been demonstrated in traditional communication media (e.g. video conferencing systems) in both listening-only and conversation tests [2, 48, 32, 17, 28], as well as for hybrid meetings [34]. While traditional communication media generally do not exhibit any three-dimensional spatiality, social Virtual Reality (VR) technologies allow users to meet and communicate in shared, immersive Virtual Environments (VEs), thus inherently supporting spatial and localizable representations of users. Although spatial audio has been shown to increase presence [12], social presence [37], and psychological immersion [30] in immersive systems, these results are from passive studies where participants do not communicate during the test. When users interact in the VE, attention

*e-mail: felix.immohr@tu-ilmenau.de
†e-mail: gareth.rendle@uni-weimar.de
‡e-mail: anton.benjamin.lammert@uni-weimar.de
§e-mail: a.neidhardt@surrey.ac.uk
¶e-mail: mzh-v@gmx.de
‖e-mail: bernd.froehlich@uni-weimar.de
**e-mail: alexander.raake@tu-ilmenau.de

may be allocated differently, modulating the importance of spatial audio. Previous work that investigated the effect of spatial auditory representations of users on communication in immersive VEs did not uncover significant effects on participant behavior or experience in an interactive dyadic conversational context [14], potentially due to a study paradigm where limited movement was required from participants users. This work aims to further investigate which communication scenarios benefit from spatial audio representations in VR. We hypothesize that in scenarios where movement is required, spatial audio will have a positive effect on user experience and behavior, due to the increased relevance of spatial hearing that arises when the location of the participants' communication partner varies.

We report on a study in which pairs of interlocutors perform a collaborative conversational task in an audiovisual VE under different *auralization* and *scene arrangement* conditions. Results are obtained through questionnaires measuring social presence and plausibility, as well as through conversational and behavioral analysis. We introduce a dedicated, spatially-aware study paradigm when compared to related work [14], in which more motion is required to complete the study task. Furthermore, we extend the evaluation methods by means of an additional plausibility questionnaire and more comprehensive behavioral analysis.

This work represents part of a wider effort to identify technical characteristics that are important for developing plausible VEs. Knowledge about the relative influence of various technical factors, such as avatar characteristics, audio rendering parameters, or visual fidelity of the VE, is a prerequisite for effective allocation of computational resources when designing and implementing VR applications. Providing plausible spatial audio for multiple sound sources and achieving spatio-temporal alignment with visual information represent significant computational and architectural challenges. Knowledge of the relevance of spatial audio to communication scenarios is therefore important when developing social VR systems.

Our research led to the following contributions:

- a novel collaborative conversation task, adapted for VR from traditional video conferencing evaluation, shown to require spatial navigation to emphasize the use of spatial hearing even in dyadic scenarios;
- a conversation study investigating the role of spatial audio in immersive communication scenarios, finding that even in collaborative tasks involving movement, no significant effect of spatial audio on dyadic communication could be shown;
- an active-listening preference test, indicating a significantly higher preference for spatial audio in direct comparison to non-spatial;
- analysis data that is made available following an open science approach[1].

2 Related Work

2.1 Audiovisual Plausibility and Presence

Slater [49] proposed that VR systems elicit realistic reactions when users experience *presence* (referred to by Slater as Place Illusion), and *plausibility* (Plausibility Illusion). Presence is regarded as the sense of 'being there' in a VE, even though one knows that one is not really there, while plausibility is the sense that the events occurring in the VE are actually occurring, even though one knows that they are not. The relationship between presence and plausibility has been further examined in later works [46, 19], and both constructs have received attention in the context of evaluating Mixed Reality (MR) experiences. *Social presence*, which is the feeling of being in a VE together with another person, is commonly used to assess the effectiveness of social MR applications [45, 4, 29]. The degree of social presence experienced by users can be considered a measure of the quality of communication media and has been shown to discriminate between different communication systems and Face-to-Face (F2F) interactions [45]. Its assessment is typically based on questionnaires, as in the semantic differential technique [45] or the Networked Minds Measure [4].

In the research field of virtual acoustics, plausibility is defined as the degree to which a virtual auditory stimulus conforms to the listener's *expectation* of an equivalent real acoustic event [5, 22], and can be assessed in listening tests through direct comparison of real and virtual stimuli [22, 27, 40]. In the VR field, plausibility relates to the expectations of the user, not only on the sensory and perceptual levels, but also on a cognitive level; if events that occur in the VE conflict with the users' expectations, which could either arise from real-world experiences or through narratives told within the VE, then plausibility is reduced [47]. Comparisons with real stimuli are avoided in the VR field due to the limited visual fidelity possible on current devices. Instead, questionnaires often specific to the given context or task have been used to measure plausibility: Hofer et al. [13] examined the effect of breaks in plausibility using a three-item questionnaire that measured alignment of the VE to users' expectations (e.g. "the rooms I walked through are very similar to rooms in real life"); and Brübach et al. [8] administered a questionnaire that was extended to 13 items, which assessed the plausibility of the behavior of scene objects with respect to both external plausibility (expectations of the user based on experiences prior to the study) and internal plausibility (whether the virtual stimuli conformed to expectations that were influenced by the narrative presented during the study).

2.2 Communication Behavior in Social VR

In addition to evaluation methods based on self-reporting, more objective means of evaluation could be derived from assessment of both verbal and non-verbal communication behavior. Verbal communication behavior can be assessed by quantifying turn-taking patterns that occur during verbal interactions through conversation analysis [38]. Based on a conversational state model of speech, metrics for dyadic interactions are derived by calculating the frequency and duration of four possible conversational states [7]: *silence*, *double talk* (when both interlocutors are speaking) and two *single talk* states (when only one interlocutor is speaking). This parametric concept has been utilized for the analysis of both audio-only [9, 32, 42] and video conferencing [44, 43, 39], showing changes in conversational structure influenced by quality degradations such as transmission delay [41].

Changes in non-verbal communication behavior can be quantified through gestural coding. Smith and Neff [50] assessed changes in non-verbal behavior in different communication modalities by classifying gestures into categories, such as referential and backchannel gestures, in a study evaluating the effect of embodiment in VR. The authors were able to show that communication and associated behavioral patterns in embodied VR was more similar to F2F communication than unembodied VR. Further aspects of non-verbal behavior include assessment of user exploration and movement behavior, as studied by Rossi et al. when investigating the influence of narrative elements in social VR [36] or exploration of dynamic point clouds [35]. Outside the context of communication scenarios, comparable means of exploration behavior based on motion and head rotation have been analyzed, for example by Robotham et al. [33] and Hendrikse et al. [11], focusing on the assessment of auditory aspects.

The effectiveness of communication and communication systems can also be evaluated by measuring performance in tasks that require communication, which can include task completion or response

[1] https://github.com/Telecommunication-Telemedia-Assessment/vr_communication_study_2

time [28, 52, 51] and accuracy or error rate [55, 51].

3 Study Description

The aim of our work is to investigate how binaural auralization affects communication in VR. The study designed to address our research question comprises an interactive audio-visual conversational communication test in VR using a test paradigm novel for VR, adapted from classical videoconferencing, cf. Sec. 3.1. The test was conducted in two strictly consecutive parts: (a) the conversation test itself followed by (b) an active listening preference test. The inclusion of a preference test was intended to determine whether participants could discriminate between auralization conditions when a direct comparison was possible without the cognitive load imposed by a conversational study task.

3.1 Conversation Test Design and Task

The conversation test followed a two-by-two factorial within-subject design, leading to four consecutive conditions as represented in Table 1. The independent variables were: the auralization method used, with binaural spatial audio (SPATIAL) being compared to a diotic audio representation (DIOTIC), and the scene arrangement. In the distributed scene arrangement (DIST), the items of interest for the task are distributed across the room, on the floor, and on a diamond-shaped table as depicted in Fig. 1, (a) and (b). In the other scene arrangement (SHAD), the items are all stacked on the table, as illustrated in Fig. 1 (c). The objects in the second scene arrangement are deliberately positioned to block the line of sight between conversation partners. It is noted that a possible break in audiovisual plausibility was expected, due to the fact that the audio shadowing effects that would be caused by the cubical items were not modelled within the audio rendering system, which is further detailed in Sect. 3.4. The two scene arrangements, DIST and SHAD, thus allow comparison both with respect to potential breaks in plausibility as well as behavioral changes due to occlusion of the conversation partner. The layout of cubic objects were the same for each participant in both scene arrangements; that is, an object on the right side of participant one's set had a corresponding object on the right of participant two, as shown in Fig. 1c. This layout enables referencing of similar task objects, while it at the same time encourages changing distance and angular differences of the positions and head orientation of the two participants in line with the goal of increasing the relevance of spatial hearing. To limit potential order and sequence effects, a balanced latin-square design was utilized.

Factor	Spatial Audio	Diotic Audio
Distributed Scene	SPATIAL_DIST	DIOTIC_DIST
Occluding Scene	SPATIAL_SHAD	DIOTIC_SHAD

Table 1: Overview of conditions and associated names based on the two independent variables of auralization method and scene arrangement.

The study task is a version of the Leavitt task, originally proposed for traditional telemeeting assessment [16], that has been adapted for a social VR context. The original Leavitt task requires study participants to identify common items in sets of colored shapes that are separately presented to each participant, like those shown in Fig. 2. While the information is naturally separated in the case of video conferencing by providing participants with sets of shapes on separate sheets of paper, VR applications typically place users in a shared space with shared visual references. If participants could see each other's sets, the need to converse is diminished. Hence, in this novel, adapted instance, shapes are presented to each participant on cubes that are distributed throughout the room. Each cube has a shape depicted on one side only, facing one participant. Participants

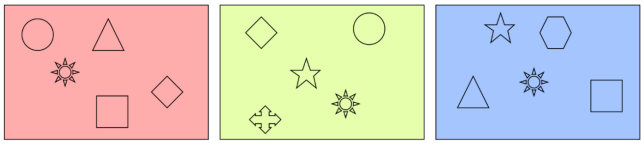


Figure 2: Exemplary sheets of the original Leavitt task for telemeeting assessment as recommended by ITU [16]. Participants must identify the common shape (in this case, the sun-like symbol).

were instructed not to cross a line dividing the room, to prevent them from solving the task by moving to view the other person's shapes, which would eliminate the need for conversation. In our test paradigm, the participants' goal is to identify when shapes on corresponding cubes were *different* by communicating with their partner. Shapes could differ in color, orientation, or form. Three pairs of cubes with differing shapes were displayed in each study trial. Participants were instructed to mark the cubes displaying shapes with differences by touching them with their virtual hand, as shown in Fig. 1(b). The trial ended once all three differing shapes were found or after a maximum of five minutes. Task completion times are evaluated in Fig. 6.

Previous related work by Immohr et al. [14] investigated the effect of spatial audio in VR communication scenarios based on a different task to elicit conversation and non-verbal communication, which did not encourage participants to move around in the VE. In this work, the novel task was chosen and adapted to encourage participant movement and hence emphasize the use of spatial hearing, even in a two-party conversation. The adaptation of this paradigm to VR maintains the need for verbal communication, and encourages non-verbal communication like the usage of gestures for shape description, or for referencing scene objects.

3.2 Participants

In a pre-test with a total of nine dyads, the set-up, testing approach and duration was verified. The pre-test participants were excluded from the main test and corresponding data was not considered in the presented analysis. In the subsequent main test, a total of sixteen dyads took part in the study, comprised of 32 participants (23 male, 9 female) aged 21-39 (M=26.94, SD=4.02). While nine dyads were mixed in gender, seven pairs consisted of two male participants. The gender distribution was a result of participant recruitment from the university body, combined with the logistical challenges of scheduling distributed multiparty VR experiments and the mitigation strategy used for non-attendance. In five dyads, participants reported that they were not familiar with each other, while in the remaining eleven pairs, familiarity as classmates, colleagues, or friends was indicated. Participants also reported prior experience in perception tests, with sixteen never, thirteen rarely (1-3 times) and three often participating in such studies before. An approval by the ethics commission of TU Ilmenau was obtained ahead of the experiment.

3.3 Procedure

Upon arrival at the test laboratories, participants were asked to fill out a consent form and a short demographic survey. This also included recording conversation partner familiarity, general perception and VR experiment experience as well as hearing abilities and diagnosed impairments. All participants were screened for visual acuity and color vision using the Snellen and Ishihara test charts. While interpupillary distance (IPD) was not formally measured prior to the test, all participants underwent the same procedure to ensure adequate stimuli presentation, during which participants were shown how to adjust the HMD's IPD settings, and were instructed to adjust the IPD until the on-screen text was clearly readable. A training phase preceded the study, after which the conversation and the active

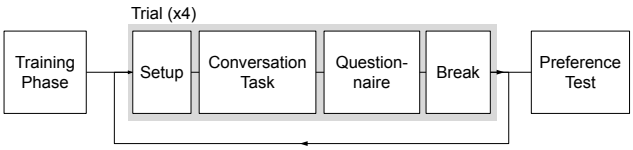


Figure 3: Experiment flow diagram.

preference tests were carried out, as shown in Fig. 3. In the training phase, participants actively explore the virtual environment and perform a simplified version of the task, gaining familiarity with each other, the virtual environment, and the devices used. The experiment took up to 90min in total. All participants received a compensation of 18€, equalling 12€ per hour.

3.3.1 Conversation Test

The conversation test consisted of four trials, with each trial corresponding to one of the four conditions presented in Table 1. Each trial started with a short setup phase, in which participants were assisted to put on the equipment and reminded of the task paradigm. Subsequently, participants performed the task (cf. Sect. 3.1). Each trial ended with a post-trial questionnaire as described in Sect. 4 that was presented in a digital form using the UNIPARK[2] survey platform, followed by a short break.

3.3.2 Active Listening Preference Test

After the conversation test, an active listening preference test was performed. Subsequent to reading through a separate set of instructions, participants were placed in an identical VE as in the conversation test, again with a different set of shapes visible to each person. Each participant was assigned one of two roles, which was visually indicated by instruction screens in VR. The order of participant roles was determined by the location, with one of two used rooms always prompting a given role first. While one participant was prompted to describe their set of shapes, the other participant was instructed to move freely, actively listen, and switch between the different auditory representations to determine which one they preferred. Immediately after listening, the subject was asked to indicate their preferred auditory condition in VR indicated by instruction screens. Participants were not informed which audio condition was spatial or diotic and were permitted to choose a neutral response to indicate no preference. Subsequently, the process was repeated with reversed roles and unassociable condition labels. After the listening preference task, subjects took part in an informal post-test interview.

3.4 Study Setup and Data Collection

The setup used in this study is illustrated in Fig. 4. The study was conducted in two similar ITU-R BT.500 [15] and ITU-T Rec. P.1301 [16] compliant laboratories. A desktop workstation in each laboratory hosted a Unity[3] application that synchronized scene state over the network using Photon Unity Networking, and transmitted speech using Photon[4] Voice 2. The experiment flow was controlled using the bmlTUX framework [3]. An identical set of hardware components, listed in Table 2, was used in each room. The VE is displayed on Head-Mounted Displays (HMDs), which track head and hand poses to animate the subjects' avatars (c.f. Fig. 1). The virtual scene is a replica of the real laboratory, as illustrated in Fig. 1. A diamond-shaped virtual table was placed in the room's center, with interlocutors placed on opposite sides of the table. The virtual room

Component	Employed Hardware
HMD	Meta Quest Pro (Oculus Link Mode)
Headphones	AKG K702
Microphone	Meta Quest Pro (integrated)
Audio Interface	MOTU M4
Desktop Computer:	
CPU	Intel Core i7-12700
Memory (RAM)	64GB
Memory (SSD)	2TB Samsung 970 EVO Plus M.2 SSD
GPU	NVIDIA GeForce RTX 3090 Ti
OS	Windows 11

Table 2: Set of hardware components used for each participant.

was bisected by a yellow line indicating the limit of the explorable area for each participant.

Questionnaire responses after each trial were recorded with the UNIPARK[2] system. In addition, the complete scene state, including head and hand orientation and position, as well as speech from each participant, were recorded for the duration of each trial with a custom analysis tool that allows recording, immersive re-exploration, and collaborative analysis of VR studies.

For the spatial audio conditions, position-dynamic binaural auralization was realized with a modified version of the open-source pyBinSim [25] renderer. A Unity Native Audio Plugin[5] was integrated for communicating user position and orientation, as well as individual source signals, to the binaural renderer via ZeroMQ[6]. For the binaural rendering, direct sound, early reflections, and late reflections are processed separately. The direct sound is dynamically synthesized in pyBinSim with energy scaling according to the inverse distance law. Here, the SADIE II Head-Related Transfer Function (HRTF) data set (subject D2 - Kemar) [1] was used alongside the mouth directivity data set (female speech) that comes with the MCRoomSim toolbox [54]. The reverberation, based on a Binaural Room Impulse Response (BRIR) measured with a KEMAR 45ba head-and-torso simulator, is not position-dynamic, since Neidhardt et al. showed that within a certain area in front of the sound source such an approach leads to an impression as plausible as an entirely measured BRIR dataset [26]. HRTF individualization for this multi-user study was not attempted to avoid the associated complex logistical and equipment challenges, since recent work points towards even generic HRTFs providing plausible reproduction with a head-tracked system [23, 26].

Diotic audio was realized with pyBinSim through a measured monaural room impulse response, in which the signal energy was adjusted to fit the binaural case, specifically the energy of the direct sound of the BRIR at the 0° azimuth reference. Distance attenuation was similarly realized based on the inverse distance law. This was done to ensure approximately equally distance-dependent loudness between the presented conditions, regardless of the distance between users, which became especially relevant for the direct active listening comparison.

4 Evaluation and Results

4.1 Questionnaires

For subjective evaluation, a two-part questionnaire was employed. The first part was based on the construct of social presence. Participants rated 16 items using a 7-point Likert scale consisting of subscales of the Networked Minds Social Presence Inventory (NM-SPI), an established social presence questionnaire [4]. These include 'Co-Presence', 'Perceived Message Understanding', and 'Mutual Assistance' [10], which were chosen as those were the subscales shown

[2] https://www.unipark.com
[3] Unity Editor 2020.3.19f
[4] https://www.photonengine.com/
[5] https://docs.unity3d.com/Manual/AudioSpatializerSDK
[6] https://zeromq.org/

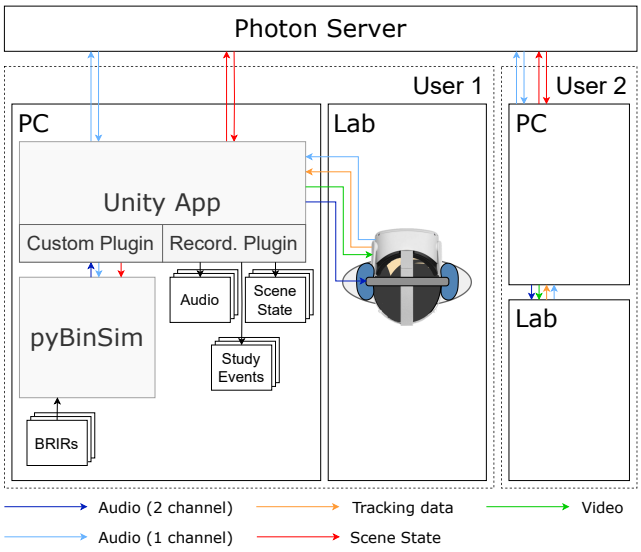


Figure 4: Structural overview of the symmetrical VR study system. The included recording plugin records scene state, audio signals, and study events from each Unity application.

to differentiate between communication media according to the original validation [4]. Mean scores of the subscales are illustrated in Fig. 5a-5c. For statistical analysis, we performed posthoc pairwise comparison employing the Wilcoxon signed-rank test subsequently to checking the normality assumption using the Shapiro-Wilk test, which was not confirmed here.

The second part of the questionnaire, on the other hand, was designed to enable a less context-dependent assessment of quality and plausibility aspects. This was focusing on stimulus presentation rather than understanding of the credibility of events in the VE. Here, participants rated 21 items, listed in Table 3, on a 7-point Likert scale. These included sub-dimensions related to task difficulty, enjoyment, interaction, audiovisual quality, and coherence as derived from the literature. Items were included from previous works [50], alongside further questions deemed relevant for the study. To avoid any specific introduction of technical terms and attention steering towards specific technical aspects, terms like 'system', 'interface', 'audio' and 'visual quality' were deliberately omitted. The questionnaire was completed by participants after each trial. Ratings for three exemplary items are shown in Fig. 5d-5f. Using the gathered results of the questionnaire, we performed an exploratory factor analysis. No significant differences were observed in the mean scores of the derived factors as a result of the conditions. The derived factor structure from the full set of questionnaire items is available in the supplementary material.

4.2 Task Performance

As a measure of task performance, the completion times for each trial are analyzed as illustrated in Fig. 6. For statistical analysis, the normality assumption was checked using the Shapiro-Wilk test, which revealed that normality cannot be assumed. The Wilcoxon signed-rank test was employed for significance testing. It becomes evident in Fig. 6a that the task exhibits a significant learning effect as the completion time decreases with an increasing number of completed trials , with the first trial being significantly shorter than the third trial ($p = 0.023$) as well as the second and third trials being shorter than the last ($p = 0.014$, $p = 0.048$). While no significant effect of the auralization method was found, tendencies suggest lower task completion times with spatial over diotic audio in both scenes.

#	Prompt
Enjoyment	
Q12	I would have liked to spend more time in the environment.
Q18	The task annoyed me.
Interaction	
Q7	I could interact with the environment.
Q11	I was able to interact fluently.
Q15	It felt intuitive to interact with the other person.
Q2	I struggled getting comfortable with the environment.
Task Difficulty	
Q8	I had difficulties solving the problem.
Q3	The task would have been easy to solve in real life.
Q16	I was easily distracted from the task.
Audiovisual Quality	
Q6	The environment was of high quality.
Q1	The conversation felt natural.
Q10	The environment sounded convincing.
Q19	The environment looked convincing.
Q14	The environment disappointed me.
Coherence	
Q4	The boxes fit into the environment.
Q9	The elements of the environment were all of the same quality.
Q17	The environment felt realistic.
Q21	It was easy to move around in the environment.
Q13	I had an influence on the environment.
Percieved Conversation Structure	
Q20	I interrupted the other person often.
Q5	I was interrupted often by the other person.

Table 3: Full list of items of the second questionnaire related to aspects of plausibility. The items were presented in order of the numbers indicated in the first column.

4.3 Exploration Behavior Analysis

The task paradigm used during the conversational test trials was designed to encourage participants to move around the VE, by requiring participants to compare and mark cubes distributed in the virtual space to complete the task. By inspecting the position distribution occupied by participants over the entire experiment, shown in Fig. 7a, it becomes evident that participants do move around the table where the cubes are placed. When compared to the heat map derived from Open Source tracking data available from previous research into the effect of spatial audio in VR communication [14] (Fig. 7b), where participants performed a different task, we can see that the task paradigm used in the current work encourages more movement. Further, the average speed as an intrinsically time-normalized measure of participant movement was analyzed. Fig. 7c - 7d show rotation and translation speed of participants heads for all trials. After the normality assumption was confirmed using the Shapiro-Wilk test, the paired samples t-test was performed. While the distributed scene arrangement naturally exhibits a significantly increased head rotation speed for both SPATIAL ($t(31) = 5.31$, $p < .0001$) and DIOTIC ($t(31) = 5.96$, $p < 0.0001$) as well as translation speed for SPATIAL ($t(31) = 4.62$, $p < .0001$) and DIOTIC ($t(31) = 2.77$, $p = .009$), a significant impact of the auralization could not be confirmed here.

In contrast to traditional communication systems, social VR supports means of non-verbal communication and behavioral patterns that are close to face-to-face interaction [50]. As the use of gestures is considered an important factor in social experiences and group interactions [18], we also inspected the mean speed of participants' hands as shown in Fig. 7e. Here, no effect was observed.

4.4 Conversation Analysis

We analyzed the conversational structure using parametric conversation analysis as described in Sect. 2.2. The analysis was applied to recordings of speech obtained from the recording tool in Unity (c.f. Sect. 3.4). As all instances of computer mediated communica-

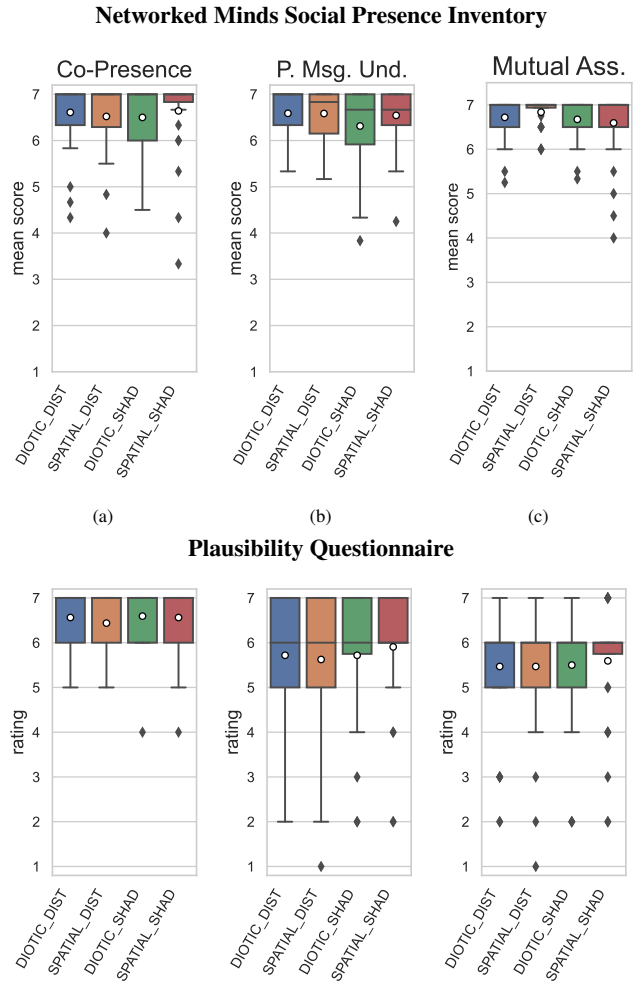


Figure 5: Questionnaire results. Subfig. 5a-5c show the mean scores of the factors Co-Presence, Percieved Message Understanding, and Mutual Assistance of the Networked Minds Social Presence Inventory. Subfig. 5d-5f show ratings for three exemplary questions relating to plausibility. All items were rated on a 7-point Likert scale.

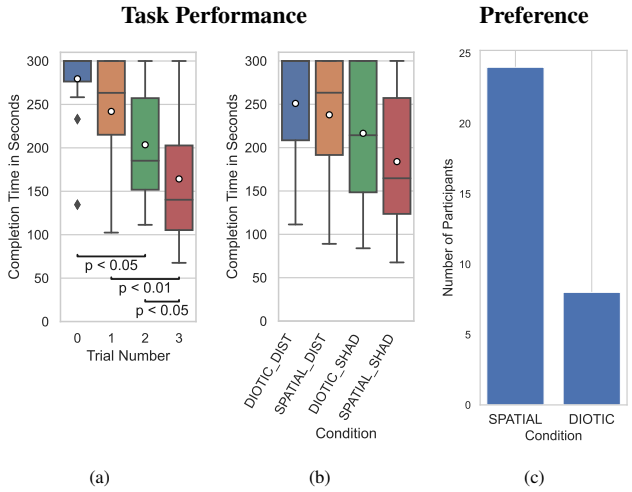


Figure 6: Participant preferences and task performance in terms of trial completion time per trial number (Subfig. 6a) and per condition (Subfig. 6b). Subfig. 6c shows preferences as reported after active listening as the final task.

tion inherently exhibit transmission delay, each utterance is delayed by the transmission channel between interlocutors. This leads to diverging sensory realities, as each participant is subject to a different conversational part of the conversation, as indicated from the surface structure. To accomodate the respective reality as percieved by each of the two participants per trial, the microphone signal of the local user as well as the spatialized receiving end of the remote user were recorded by each client. Subsequent to normalization of signal levels, a signal-based Voice Activity Detection (VAD) was performed. As recommended in [7], missing voice activity was only classified as silence if longer than 200ms, while talkspurts shorter than 15ms were treated as silence. Based on the derived active speech patterns, a per-sample representation of states is computed based on a conversational state model (c.f. Sect. 2.2). From this representation, further metrics are derived, including state probabilities as ratios of the conversation spent in a given state, as well as state sojourn times, which are the average duration a given state has been maintained. On the computed metrics, the normality assumption was checked individually employing the Shapiro-Wilk test. Accordingly, post-hoc pairwise t-tests (parametric) or Wilcoxon signed-rank tests (non-parametric) were performed with Bonferroni correction. Results are illustrated in Fig. 8. While a significant effect of the auralization has not been confirmed, we found significantly higher probability ($W = 108$, $p = .017$) and sojourn time ($W = 107$, $p = .016$) of the 'mutual silence' state under the distributed scene arrangement condition with spatial audio. The same can be observed between DIOTIC_SHAD and SPATIAL_DIST (state probability: $W = 109$, $p = .018$; sojourn time: $W = 96$, $p = .007$) as well as SPATIAL_SHAD and DIOTIC_DIST (state probability: $W = 71$, $p < .001$; sojourn time: $W = 121$, $p = .039$). This is potentially corresponding to increased exploratory movement and longer task completion times also observed for that scene arrangement condition.

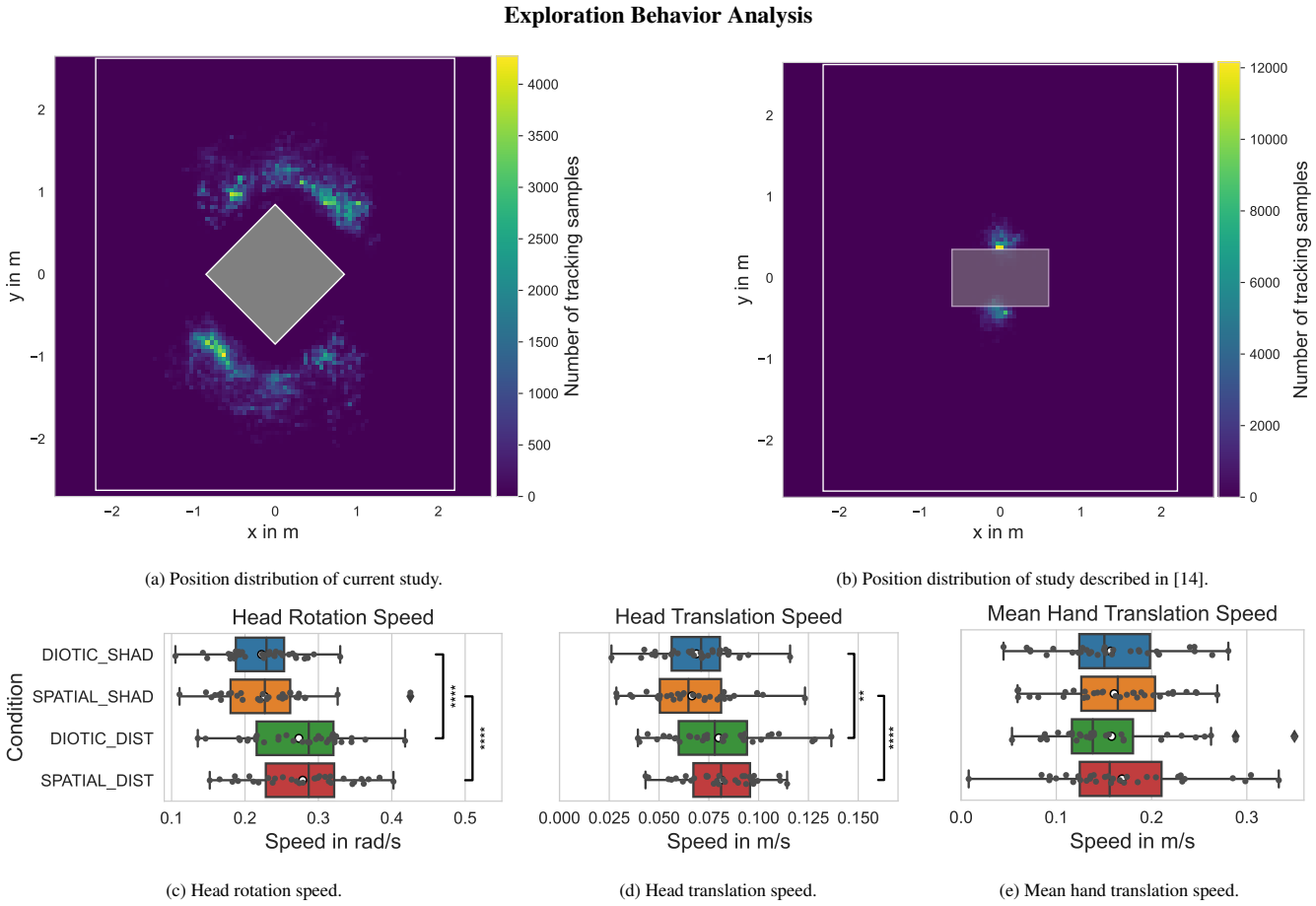
4.5 Active Listening Preference

All participants indicated their preferred auralization method after active listening as described in Sec. 3.3.2. Results are depicted in Fig. 6. While 24 participants preferred spatial audio (SPATIAL), 8 participants preferred the non-spatial representation (DIOTIC). No participants chose the neutral response.

5 Discussion and Limitations

The study presented in this work aims to investigate the effect of auralization techniques on user experience and behavior in VR communication scenarios. While previous work had shown little impact of spatial audio in such dyadic scenarios [14], we designed a task that would encourage more movement to increase the significance of spatial hearing. By analyzing tracking data from this study and previous work, it can be seen that the task succeeded in requiring participants to move.

We hypothesized that when participants moved more while conversing in the VE, spatial audio would be more noticeable for participants. However, responses to questionnaires investigating social presence and plausibility did not show significant differences caused by varying the auralization condition. Analysis of recorded scene state data also did not confirm a significant effect of spatial audio on behavioral measures. However, the scene arrangement as a context factor showed an effect on both the verbal behavior, specifically the conversation structure, and non-verbal behavioral metrics. In



(a) Position distribution of current study.

(b) Position distribution of study described in [14].

(c) Head rotation speed.

(d) Head translation speed.

(e) Mean hand translation speed.

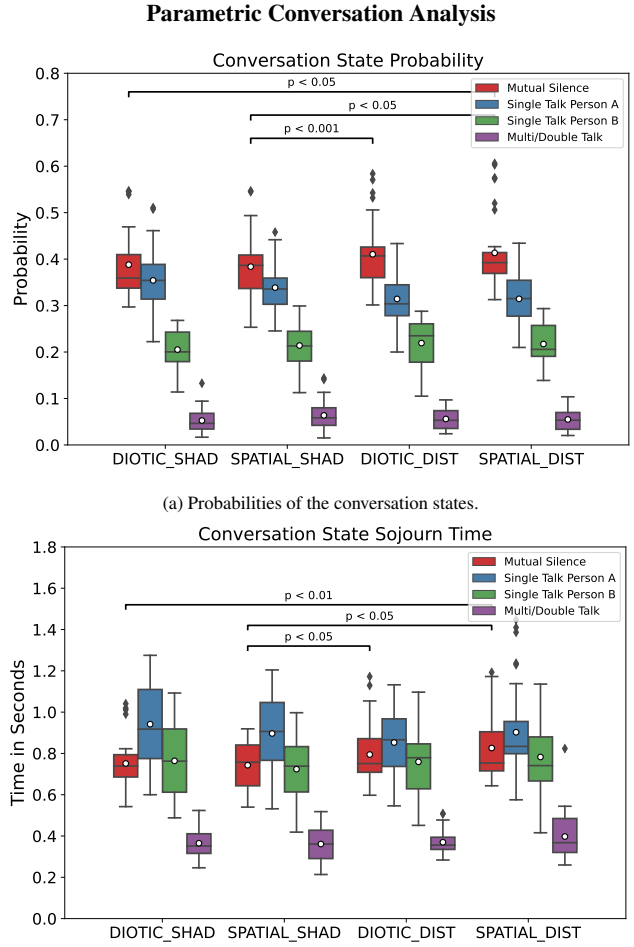
Figure 7: Exploration behavior, based on position distribution and movement speed. Subfig. 7a and 7b show user positions during the test as 2D histogram of all uniformly sampled tracking points of all trials. To match the sampling rate of tracking data from the presented study, the tracking data from [14] was downsampled to 25Hz. White lines enclose the movable area. Gray rectangles represent the virtual table carrying (parts) of the scene objects relevant to solve the respective task. Subfig. 7c-7e show movement speed of participants.

contrast, direct comparison of the auralization conditions showed a significant preference for the spatial audio condition. While the discrepancy between results from the direct comparison in the preference test and the indirect measurements from the conversation test is somewhat surprising, there are some reasons that may account for this.

Evaluation Methods One potential explanation relates to the contrasting evaluation methods. The questionnaires administered after each trial did not directly ask participants to evaluate audio quality aspects, such that participants were not prompted to focus their attention on audio stimuli. This aimed to avoid distracting participants from performing the conversation task in a natural manner or leading them to more critically judge the quality of certain modalities. Instead, questions related to the interaction with the other person, and judgments on aspects of the environment were prompted. The indirect and holistic nature of the questions and constructs like social presence and plausibility could mean that they were not suitable for detecting how participants perceived the audio presentation, although psychoacoustic changes were audible as confirmed in the preference test. Although current state-of-the-art methods used in relevant studies [50, 21] for assessing social presence often involve questionnaires with more than 20 items [4, 20, 24], such high number of ratings might limit the accuracy of measurement due to questionnaire fatigue, especially when repeated in within-subject designs. However, prior work on the effect of spatial audio in video conferencing did detect effects from similarly indirect questionnaire items regarding the perceived degree of interactivity and the feeling of sharing a space with collaborators [28]. To augment self-reporting based methods, recent work aims to gain a more robust understanding of objective behavioral indicators (c.f. Sect. 2.2) and motivated this work as well.

Furthermore, listener expertise plays an important role in listening test performance of subjects, for example in the context of room acoustics assessment [53]. With the majority of participants in our study reporting to be naive in this type of perception tests, the ability to perceive differences between auralization methods in a conversational setting might be limited by listening expertise and experience.

Technical Limitations Further, the presented situation only exhibited limited acoustic scene complexity, with just two active speakers in controlled lab environments. The low-reverberance nature of the used rooms prevents the introduction of unwanted artifacts to the transmission path, for example, reverberation and room characteristics captured by the microphones. Although suitable and recommended [16] for communication system assessment, this environment does not exhibit realistic or acoustically complex listening situations, particularly when the task is performed in a dyadic constellation. The real and virtual representation of the VE were kept similar in the visual as well as auditory domain. This was done to prevent auditory plausibility violations when the head-



(a) Probabilities of the conversation states.

(b) Mean sojourn times of the conversation states.

Figure 8: Parametric conversation analysis results. Subfig. 8a shows state probabilites of the four conversation states (c.f. Sect. 2), providing insights into the conversation surface structure. In this representation, the more active speaker in each recorded conversation is associated as person A. Subfig. 8b illustrates the mean sojourn times of the same states (the average duration a given state has been maintained).

set was removed between trials (c.f. Fig. 3), since the subjective evaluation methods were centered around the related constructs of social presence and plausibility. Non-verbal communication plays an important role for both the interaction itself as well as behavioral analysis for evaluation of technical factors (c.f. Sect. 2.2). While the user representation enabled gesturing, for example pointing towards items or describing a given shape, the full set of potential expressions was limited by the system only providing three tracked six-degrees-of-freedom controllers per user. Although being important aspects, facial tracking, posture or finger movement were hence not represented, potentially limiting the relevance of movements both for participants and as an evaluation metric.

Task and and Context Limitations The discrepancy between results from the conversation test and preference test may also relate to the respective tasks involved. During the conversation test, participants' attention is focused on completing the study task, instead of evaluating audio quality, and focussing especially on the visual modality (shape recognition). It may be that the cognitive load of the study task detracts from the ability to perceive differences in auditory stimuli. Alternatively, it may be that, although participants move around the VE, the spatial information provided by the auditory cues is not important enough for the task at hand. In this dyadic communication scenario, only one other person can be speaking, meaning that audio cues are not used for differentiating between speakers. Although spatial audio cues could give information about the location of the other speaker, the specific context and task may not require movement fast enough or often enough for that information to be significantly helpful. Literature investigating brain activity during spatial hearing tasks has found that brain activity in single-stimuli hearing tasks was not modulated as a function of stimulus location [56].

In this sense, our proposed study paradigm novel for VR may require further adjustment to generalize well to communication scenarios. In future work investigating technical factors in social VR, such as spatial audio, a task should be carefully chosen to avoid task complexity, which may divert attention from the technical factor in question, while encouraging actions that emphasize the factor's relevance. With respect to our study, a lower task-induced cognitive load would be achieved by simplifying the number of task objects (e.g. here number of presented shapes) or by selecting a task paradigm requiring less visual information. The relevance of spatial audio for communication in VR can be emphasized by increasing acoustic scene complexity and further encouraging verbal discussion. Extending the paradigm to include more interlocutors, meaning that spatial auditory cues become relevant for identifying the current speaker, would enable better evaluation of the role of spatial audio in VR communication scenarios. Work on spatial attention allocation has shown that processing of spatial stimuli is transient when other features can be used to distinguish audio streams [6], meaning that the adjustment of the task to involve three participants may only increase the relevance of spatial audio when participants have voices with similar tonal qualities.

6 Conclusion and Outlook

In this work, we investigate the impact of auralization method and scene arrangement on communication behavior and plausibility perception in social audiovisual VR. We have proposed a novel study paradigm, adapted from an existing conversation task, that encourages verbal and non-verbal communication between participants, as well as prompting users to move within the VE. While participants indicated a clear preference for spatial audio in a direct listening comparison between auralization conditions, no significant differences were revealed through analysis of questionnaire responses or behavioral data in the conversational context. In future work, we intend to extend this method towards higher communication complexities and hence more realistic scenarios in relation to both more life-like interactions as well as to current developments in the application scope of social VR. Specifically, this could include the involvement of additional simultaneous participants or acoustically complex scenes involving further sound sources and/or reverberation. Furthermore, the scenario could be compared to a real-world face-to-face interaction as an explicit reference. If the auralization method is found to affect user behavior and experience in VR, then investigating its relative influence with respect to other environment factors, such as avatar and environment appearance, becomes relevant.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG) under the project "Audiovisual Plausibility and Experience in Multi-Party Mixed Reality" (APlausE-MR) (ID: 444831328) as part of the DFG Priority Program SPP2236-AUDICTIVE. Further, this work has partially been supported by the DFG project "Interconnected Lab for MEdia Technology Analytics" (ILMETA) (ID: 438822823). The study has been carried out at TU Ilmenau. The authors would like to thank the study participants.

REFERENCES

[1] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney. A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database. *Applied Sciences*, 8(11), 2018. doi: 10.3390/app8112029

[2] J. J. Baldis. Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pp. 166–173. Association for Computing Machinery, New York, NY, USA, 2001. doi: 10.1145/365024.365092

[3] A. O. Bebko and N. F. Troje. Bmltux: Design and control of experiments in virtual reality and beyond, 2020. doi: 10.31234/osf.io/arvkf

[4] F. Biocca, C. Harms, and J. Gregg. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. *4th annual International Workshop on Presence, Philadelphia*, 01 2001.

[5] J. Blauert and U. Jekosch. Concepts behind sound quality: Some basic considerations. In *Proc. Internoise*, pp. 72–79, 2003.

[6] L. M. Bonacci, S. Bressler, and B. G. Shinn-Cunningham. Nonspatial features reduce the reliance on sustained spatial auditory attention. *Ear and hearing*, 41(6):1635–1647, 2020. doi: 10.1097/AUD.0000000000000879

[7] P. T. Brady. A statistical analysis of on-off patterns in 16 conversations. *Bell System Technical Journal*, 47(1):73–91, 1968.

[8] L. Brübach, F. Westermeier, C. Wienrich, and M. E. Latoschik. Breaking plausibility without breaking presence - evidence for the multi-layer nature of plausibility. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2267–2276, 2022. doi: 10.1109/TVCG.2022.3150496

[9] S. Egger, R. Schatz, and S. Scherer. It takes two to tango-assessing the impact of delay on conversational interactivity on perceived speech quality. In *11th Annual Conf. of the Int. Speech Communication Association*, 2010.

[10] C. Harms and F. Biocca. Internal consistency and reliability of the networked minds measure of social presence. In *Seventh annual international workshop: Presence*, vol. 2004. Universidad Politecnica de Valencia Valencia, Spain, 2004.

[11] M. M. E. Hendrikse, G. Llorach, V. Hohmann, and G. Grimm. Movement and gaze behavior in virtual audiovisual listening environments resembling everyday life. *Trends in Hearing*, 23:2331216519872362, 2019. PMID: 32516060. doi: 10.1177/2331216519872362

[12] C. Hendrix and W. Barfield. The Sense of Presence within Auditory Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 5(3):290–301, 1996. doi: 10.1162/pres.1996.5.3.290

[13] M. Hofer, T. Hartmann, A. Eden, R. Ratan, and L. Hahn. The role of plausibility in the experience of spatial presence in virtual environments. *Frontiers in Virtual Reality*, 1, 2020. doi: 10.3389/frvir.2020.00002

[14] F. Immohr, G. Rendle, A. Neidhardt, S. Göring, R. R. R. Rao, S. A. Arboleda, B. Froehlich, and A. Raake. Proof-of-concept study to evaluate the impact of spatial audio on social presence and user behavior in multi-modal VR communication. In *Proc. of the ACM Int. Conf. on Interactive Media Experiences (IMX)*, 2023.

[15] ITU-R Rec. BT.500-14. Methodologies for the subjective assessment of the quality of television images, 2019.

[16] ITU-T Rec. P.1301. Subjective quality evaluation of audio and audiovisual multiparty telemeetings, 2017.

[17] Janto Skowronek and Alexander Raake. Assessment of Cognitive Load, Speech Communication Quality and Quality of Experience for spatial and non-spatial audio conferencing calls. *Speech Communication*, 66:154–175, 2015. doi: 10.1016/j.specom.2014.10.003

[18] N. C. Krämer, B. Tietz, and G. Bente. Effects of embodied interface agents and their gestural activity. In G. Goos, J. Hartmanis, J. van Leeuwen, T. Rist, R. S. Aylett, D. Ballin, and J. Rickel, eds., *Intelligent Virtual Agents*, vol. 2792 of *Lecture Notes in Computer Science*, pp. 292–300. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003. doi: 10.1007/978-3-540-39396-2_49

[19] M. E. Latoschik and C. Wienrich. Congruence and Plausibility, Not Presence: Pivotal Conditions for XR Experiences and Effects, a Novel Approach. *Frontiers in Virtual Reality*, 3, jun 2022. doi: 10.3389/frvir.2022.694433

[20] J. Li, Y. Kong, T. Röggla, F. De Simone, S. Ananthanarayan, H. De Ridder, A. El Ali, and P. Cesar. Measuring and understanding photo sharing experiences in social virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019.

[21] J. Li, S. Subramanyam, J. Jansen, Y. Mei, I. Reimat, K. Ławicka, and P. Cesar. Evaluating the user experience of a photorealistic social vr movie. In *Int. Symp. on Mixed and Augmented Reality (ISMAR)*, pp. 284–293. IEEE, 2021.

[22] A. Lindau and S. Weinzierl. Assessing the plausibility of virtual acoustic environments. *Acta Acustica united with Acustica*, 98:804–810, 2012. doi: 10.3389/fnins.2013.12345

[23] T. Lübeck and C. Pörschmann. Evaluating the plausibility of non-individual head-related transfer functions in anechoic conditions. In *10th Convention of the European Acoustics Association: Forum Acusticum 2023*, 2023.

[24] G. Makransky, L. Lilleholt, and A. Aaby. Development and validation of the multimodal presence scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior*, 72:276–285, jul 2017. doi: 10.1016/j.chb.2017.02.066

[25] A. Neidhardt, F. Klein, N. Knoop, and T. Köllmer. Flexible python tool for dynamic binaural synthesis applications. In *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[26] A. Neidhardt, A. Tommy, and A. Pereppadan. Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets. In *144h Int. AES Convention, Milan, Italy*, 2018.

[27] A. Neidhardt and A. M. Zerlik. The availability of a hidden real reference affects the plausibility of position-dynamic auditory ar. *Front. Virtual Real., 06 September*, 2021. doi: 10.3389/frvir.2021.678875

[28] K. Nowak, L. Tankelevitch, J. Tang, and S. Rintel. Hear We Are: Spatial Audio Benefits Perceptions of Turn-Taking and Social Presence in Video Meetings. In S. Boll, A. Cox, T. Ludwig, and M. E. Cecchinato, eds., *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, pp. 1–10. ACM, New York, NY, USA, 2023. doi: 10.1145/3596671.3598578

[29] C. S. Oh, J. N. Bailenson, and G. F. Welch. A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI*, p. 114, 2018.

[30] T. Potter, Z. Cvetkovic, and E. de SENA. On the Relative Importance of Visual and Spatial Audio Rendering on VR Immersion. *Front. Signal Process. - Audio and Acoustic Signal Processing*, 2022.

[31] V. P. Pulkki, Matti Karjalainen. *Communication Acoustics*. John Wiley & Sons, Jan. 2015.

[32] A. Raake, J. Ahrens, M. Geier, and C. Schlegel. Listening and conversational quality of spatial audio conferencing. In *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*, Oct 2010.

[33] T. Robotham, A. Singla, A. Raake, O. S. Rummukainen, and E. A. P. Habets. Influence of multi-modal interactive formats on subjective audio quality and exploration behavior. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, IMX '23, p. 115–128. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3573381.3596155

[34] L. Rosset, H. Alavi, S. Zhong, and D. Lalanne. Already it was hard to tell who's speaking over there, and now face masks! can binaural audio help remote participation in hybrid meetings? In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411763.3451802

[35] S. Rossi, I. Viola, and P. Cesar. Behavioural analysis in a 6-dof vr system: Influence of content, quality and user disposition. In *Proceedings of the 1st Workshop on Interactive eXtended Reality*, pp. 3–10, 2022.

[36] S. Rossi, I. Viola, J. Jansen, S. Subramanyam, L. Toni, and P. Cesar. Influence of narrative elements on user behaviour in photorealistic social vr. In *Proceedings of the International Workshop on Immersive Mixed and Virtual Environment Systems (MMVE'21)*, pp. 1–7, 2021.

[37] S. Roßkopf, L. Kroczek, F. Stärz, M. Blau, S. van de Par, and A. Mühlberger. Comparable Sound Source Localization Of Plausible Auralizations And Real Sound Sources Evaluated In A

Naturalistic Eye-Tracking Task In Virtual Reality. Preprint doi: https://doi.org/10.31234/osf.io/vf5py, 2023.

[38] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pp. 7–55. Elsevier, 1978.

[39] M. Schmitt, J. Redi, D. Bulterman, and P. S. Cesar. Towards individual qoe for multiparty videoconferencing. *Trans. on Multimedia*, 20(7):1781–1795, 2017.

[40] C. Schneiderwind and A. Neidhardt. Discriminability of concurrent virtual and real sound sources in an augmented audio scenario. In *152nd AES Convention, paper 10604, The Hague, The Netherlands / Online*, 2022.

[41] K. Schoenenberg. *The quality of mediated-conversations under transmission delay*. PhD thesis, Technische Universität Berlin, 2016. doi: 10.14279/DEPOSITONCE-4990

[42] K. Schoenenberg, A. Raake, S. Egger, and R. Schatz. On interaction behaviour in telephone conversations under transmission delay. *Speech Communication*, 63-64:1–14, sep 2014. doi: 10.1016/j. specom.2014.04.005

[43] K. Schoenenberg, A. Raake, and P. Lebreton. Conversational quality and visual interaction of video-telephony under synchronous and asynchronous transmission delay. In *6th Int. Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 31–36. IEEE, 2014.

[44] A. J. Sellen. Remote conversations: The effects of mediating talk with technology. *Human–Computer Interaction*, 10(4):401–444, 1995. doi: 10.1207/s15327051hci1004_2

[45] J. Short, E. Williams, and B. Christie. *The social psychology of telecommunications*. Toronto; London; New York: Wiley, 1976.

[46] R. Skarbez, F. P. Brooks, Jr., and M. C. Whitton. A survey of presence and related concepts. *ACM Comput. Surv.*, 50(6), Nov. 2017. doi: 10.1145/3134301

[47] R. T. Skarbez. *Plausibility illusion in virtual environments*. PhD thesis, University of North Carolina at Chapel Hill Graduate School, 2016.

[48] J. Skowronek, A. Raake, G. H. Berndtsson, O. S. Rummukainen, P. Usai, S. N. B. Gunkel, M. Johanson, E. A. P. Habets, L. Malfait, D. Lindero, and A. Toet. Quality of Experience in Telemeetings and Videoconferencing: A Comprehensive Survey. *IEEE Access*, 10:63885–63931, 2022. doi: 10.1109/ACCESS.2022.3176369

[49] M. Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364:3549–3557, 2009.

[50] H. J. Smith and M. Neff. Communication behavior in embodied virtual reality. In CHI, ed., *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–12. ACM, New York, NY, 2018. doi: 10.1145/3173574.3173863

[51] S. Uhrig, A. Perkis, S. Möller, U. P. Svensson, and D. M. Behne. Effects of spatial speech presentation on listener response strategy for talker-identification. *Frontiers in Neuroscience*, 15, 2022. doi: 10.3389/fnins.2021.730744

[52] S. Van Damme, F. Velde, M. Sameri, F. De Turck, and M. Torres Vega. A haptic-enabled, distributed and networked immersive system for multi-user collaborative virtual reality. In *Second International Workshop on Interactive Extended Reality (IXR)*. Ottawa, Canada, 09 2023. doi: 10.1145/3607546.3616804

[53] M. von Berg, J. Steffens, S. Weinzierl, and D. Müllensiefen. Assessing room acoustic listening expertise. *The Journal of the Acoustical Society of America*, 150(4):2539–2548, 10 2021. doi: 10.1121/10.0006574

[54] A. Wabnitz, N. Epain, C. Jin, and A. van Schaik. Room acoustics simulation for multichannel microphone arrays. In *Int. Symp. on Room Acoustics, ISRA, Melbourne, Australia*, 2010.

[55] Y. Wu, Y. Wang, S. Jung, S. Hoermann, and R. W. Lindeman. Using a fully expressive avatar to collaborate in virtual reality: Evaluation of task performance, presence, and attraction. *Frontiers in Virtual Reality*, 2, apr 2021. doi: 10.3389/frvir.2021.641296

[56] R. J. Zatorre, M. Bouffard, P. Ahad, and P. Belin. Where is 'where' in the human auditory cortex? *Nature Neuroscience*, 5(9):905–909, 2002. doi: 10.1038/nn904