Subjective Evaluation of the Impact of Spatial Audio on Triadic Communication in Virtual Reality

Felix Immohr*, Gareth Rendle†, Christian Kehling‡, Anton Lammert†, Steve Göring*,
Bernd Froehlich†, Alexander Raake*
*Audiovisual Technology Group; ‡Electronic Media Technology Group, Technische Universität Ilmenau, Germany
Email: {felix.immohr, steve.goering, alexander.raake, christian.kehling}@tu-ilmenau.de
†Virtual Reality and Virtualization, Bauhaus-Universität Weimar, Germany
Email: {gareth.rendle, anton.lammert, bernd.froehlich}@uni-weimar.de

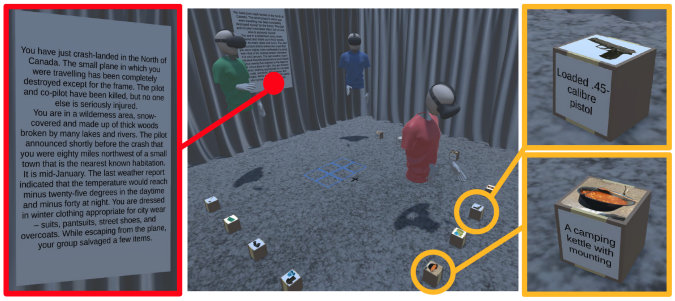
***Abstract*—Virtual Reality (VR) enables users to meet, converse, and collaborate in shared virtual environments. For such communication systems, many system factors can affect user experience and perception. To effectively allocate system resources, understanding of the relative influence of such factors is required. One important factor is a spatial auralization, which has been shown to elevate users' experience in traditional and single-user VR systems. However, its effect in multi-party social VR has not been fully investigated. In this work, we conducted a study assessing the effect of spatial audio on audiovisual plausibility and presence perception in a three-user interactive communication scenario. Triads of participants perform a collaborative conversation task under three conditions: a VR condition with binaural spatial audio, a VR condition with simple diotic audio, and a real-world reference condition. This paper presents the results of the study based on questionnaire-based evaluation.**

***Index Terms*—Virtual Reality, Subjective Evaluation, Spatial Audio, Plausibility, Social Presence, Conversation User Study**

I. Introduction

Emerging immersive communication systems, such as social VR, allow users to meet and work together in shared Virtual Environments (VEs), enabling close-to-real-life interactions beyond the limitations of traditional systems. With many system factors affecting experience and perception in this context [1], a realistic auditory representation is one important factor that is typically not considered or evaluated in multi-user studies [2, 3, 4]. A positive effect of spatial audio on a.o. Quality of Experience (QoE), psychological immersion and cognitive load has been shown in listening situations [5, 6], traditional media [7, 8] and single-user VR [9, 10, 11]. For social VR, relevant quality indicators also include social presence (c.f., e.g. [2, 3, 4]) and (audiovisual) plausibility [12], typically assessed with direct questionnaire-based methods like the Networked Minds Social Presence Inventory (NM-SPI) [13]. Investigating communication and the impact of the auralization method in multi-party VEs depends on the employed scenario and context, with increasing effect for higher scene complexities [6]. Therefore, in this work we investigate the effect of spatial audio on social presence and audiovisual plausibility in VR with three interlocutors and compare it to a real-world reference. Since dyadic studies in spaces of low acoustic complexity show little effect [14], we increase complexity to a triadic scenario. To realize the reference condition, the study was conducted in low-reverberant lab rooms and matching VE to avoid bias through different audiovisual appearance and salience in diverging virtual and real environments.

This work was funded by Deutsche Forschungsgemeinschaft (DFG) with the project "Audiovisual Plausibility and Experience in Multi-Party Mixed Reality" (444831328) as part of DFG Prio. Progr. SPP2236-AUDICTIVE. This work has partially been supported by DFG ILMETA (438822823).



(a) Scenario desc. (b) Survival task scene (c) Survival items

Fig. 1: Survival game task implementation in triadic VR.

Our research led to the following contributions:

- a novel realization of an established conversation task, adapted for multi-party VR and real-world replication;
- a conversation study investigating the role of spatial audio in immersive triadic communication scenarios with comparison to a real-world interaction.

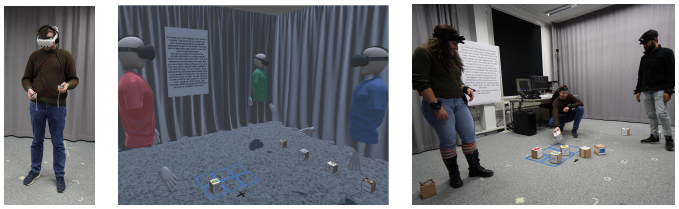
II. Study Description

Since this work aims to evaluate the influence of spatial audio reproduction on communication in three-user VR scenarios, we designed a study based on an interactive conversation task that asked participants to rate communication experience with and without spatial audio.

A. Study Design and Task

The study undertaken by each triad consisted of three trials in a within-subject design: a VR condition with diotic, non-spatial audio (DIOTIC); a VR condition with binaural spatial audio (SPATIAL); and a real-world interaction (REAL).

The employed task was adapted for room-scale VR from the Survival Task in ITU-T Rec. P.1301 Appx.VI [15, 16] originally designed for multi-party assessment of traditional



(a) Setup (b) VR survival scene (c) REAL survival scene

Fig. 2: Survival game study scene and setup in VR and REAL.

telemeeting systems. The participants' goal is to select six out of twelve presented items that would best help the group to survive in the given situation (e.g. being lost in the desert after a plane crash). The task ended when six items were placed in the marked area, or when 6 minutes elapsed. Instead of presenting information separately to participants on paper, this spatialized VR version presents the survival items, the scenario description and a marked space for the selected objects inside the VE, as shown in Fig. 1. The survival items are represented as manipulable boxes (11×11x10cm), with an image of the survival item on the top and a short item description on the front. Scenario descriptions and items of the three scenarios *desert*, *winter* and *sea* were used unaltered from the recommendation. The item illustrations were updated with higher resolution imagery to better fit the VE used here. The boxes were distributed on the floor facing inward, towards the starting positions of participants, requiring spatial and rotational exploration. In the REAL condition, the scenario description was printed on a large sheet and the items were crafted out of cardboard boxes matching the VR counterparts in appearance, size and position (c.f. Fig 2).

A greco-latin square design was employed to counterbalance the ordering of conditions and survival scenarios.

B. Procedure

After arrival, participants were asked to fill out a consent form and a short demographic survey, which included questions on conversation partner familiarity, general perception test experience, and hearing abilities. A Snellen test chart was used for visual acuity screening. Each participant underwent an Interpupillary Distance adjustment procedure to ensure adequate stimuli presentation on their Head-Mounted Display (HMD). A training phase preceded the study conditions, in which participants performed a simplified version of the task to gain familiarity with each other, the virtual scene, and the equipment. After each condition, participants were asked to fill out a digital questionnaire using the UNIPARK [17] platform, followed by a short break. The experiment took up to 90 min in total and participation was compensated with 18€.

C. Study Setup and Data Collection

The system used in the study is illustrated in Fig. 3 with the symmetrically employed hardware components listed in Tab. I. The system is driven by a Unity application that shares audio and scene state over the network through Photon Unity Networking and Photon Voice 2 [18]. To control the experimental

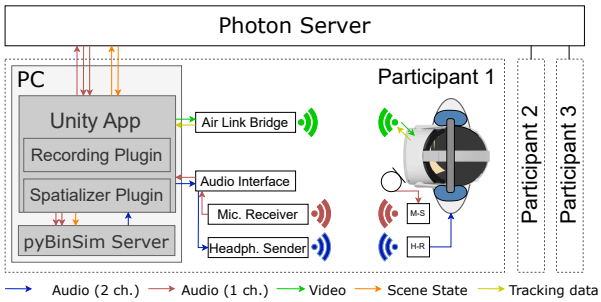


Fig. 3: Illustration of the symmetrical VR setup.

Component	Employed Hardware
HMD	Meta Quest 3 (Air Link Mode)
Air-Link Bridge	D-Link DWA-F18
Headset	Beyerdynamic DT290
Audio Interface	MOTU M4
Wireless audio (analog)	Sennheiser ew IEM G4 & Shure QLX-D
Tracking (REAL)	HTC Vive Tracker 3.0 with SteamVR Base Station 2.0
Microphone (REAL)	Shure MX150 lavalier mic
Desktop Computer:	i7-13700K, 64GB RAM, NVidia RTX 4080, Win 11

TABLE I: Hardware components used for each participant.

flow, the bmlTUX framework [19] was used. Participants wore an HMD to view and interact with the Unity application, which rendered the VE. The avatars were animated by the tracked positions of the HMD and the controllers.

Position-dynamic binaural audio is realized with an extended version of the open-source pyBinSim [20] renderer, which receives position information and remote users' transmitted microphone signals from instances of a Unity Audio Spatializer plugin [21], before returning the processed audio. While the direct sound is dynamically synthesized with the SADIE II Head-Related Transfer Function (HRTF) database (subject D2 - Kemar) [22] and the speech directivity dataset (female speech) provided in [23], the late reverberation is based on a Binaural Room Impulse Response (BRIR) set measured with a KEMAR 45ba head-and-torso simulator in a room similar to the used laboratory. Direct sound energy is scaled using the inverse distance law. Reverberation was not position-dependent, since it was shown that similar approaches lead to an equally plausible impression as an entirely measured BRIR dataset, if close to a frontal sound source [24]. The DIOTIC condition was realized directly in Unity without rendering of distance attenuation or room characteristics. The loudness was adjusted so that the average sound pressure level at 1.5m distance is equal between the presented conditions.

For recording of individual speech and scene states, including tracking data and study events, a Unity-based recording plugin was implemented. In the REAL condition, the Unity system was leveraged for consistent control and recording of the experiment. Instead of using HMDs and headsets for recording of speech and tracking data, each participant wore a lavalier microphone and HTC Vive tracker on hands and head.

III. Evaluation and Discussion

A total of 22 triads completed the study, equalling 66 participants (42 male, 24 female, none diverse) aged 21-38 years

(μ=26.79, σ=3.72) and recruited from the university body. While 16 triads were mixed in gender, five groups consisted of three male and one group of three female participants. In nine triads, participants reported no familiarity, and in four groups all participants reported familiarity (acquaintance or higher). In the remaining nine triads some degree of familiarity was indicated between two persons, but one participant reported no familiarity. An approval by the ethics commission of TU Ilmenau was obtained ahead of the experiment.

After each trial, participants were asked to respond to a series of questionnaires with 29 items in total. The first item queried Overall Experience, which was rated on a five-point absolute category rating scale. No significant effect was observed. This was followed by 16 items from the NM-SPI [25, 13] (for Co-Presence, Perceived Message Understanding and Mutual Assistance subscales), which were rated on a seven-point Likert scale. After rejecting the normality assumption with the Shapiro-Wilk test, the Wilcoxon Signed-Rank test was performed. The results are presented in Fig. 4.

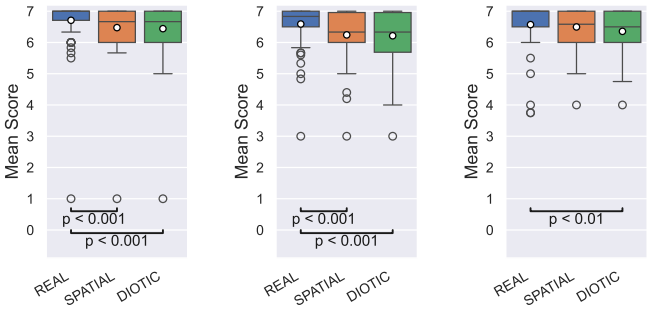
A further questionnaire with twelve items evaluated aspects of audiovisual plausibility, like coherence, interaction and quality (rated on a 1-7 Likert scale). With the normality assumption not confirmed, the Mann-Whitney U test was performed on the individual items. Ratings of three exemplary items are depicted in Fig. 5. While there are tendencies for increased social presence and plausibility aspects with spatial over diotic audio in this communication context, a significant effect is only found in comparison to the real-world condition.

After the experiment, participants were asked to rank the trials. The most preferred and least preferred conditions and scenarios are shown in Fig. 6. While REAL was strongly preferred, DIOTIC was least preferred by a smaller margin.

While it is expected that the REAL condition is preferred to the VR conditions in both the direct ranking and the questionnaire results, the lack of evidence that the spatial auralization condition is preferred over the diotic condition is surprising. One factor that may limit the impact of spatial auralization is the sensitivity of the evaluation methods to changes in audio presentation, as naive listeners rated holistic multi-modal metrics, without their attention being directed towards specific system modalities. Differences may be revealed through objective analysis of behavioral and physiological measures. Another factor could be the low acoustic scene and communication complexity; since the study was performed in quiet, low-reverberance lab rooms and in small groups, spatial auditory information might only be of limited importance. Furthermore, the conversational task might require a high cognitive load which could potentially shift attentional allocation away from specific system aspects.

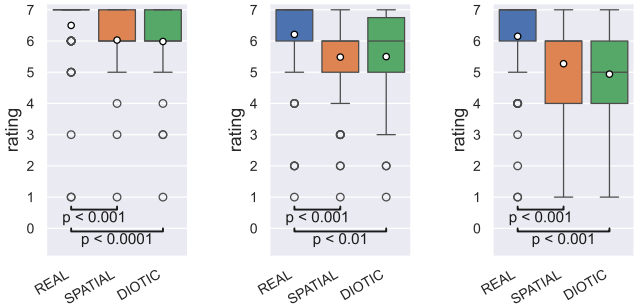
IV. Conclusion and Outlook

In this work, we assessed the impact of spatial audio in VR communication, with reference to a real-world interaction. A study with 66 participants was designed and conducted, in which triads performed a spatialized version of the survival task. Social presence, aspects of plausibility, and participant



(a) Co-Presence (b) Perc. Msg. Und. (c) Mut. Assistance

Fig. 4: Social presence results of three NM-SPI subscales.

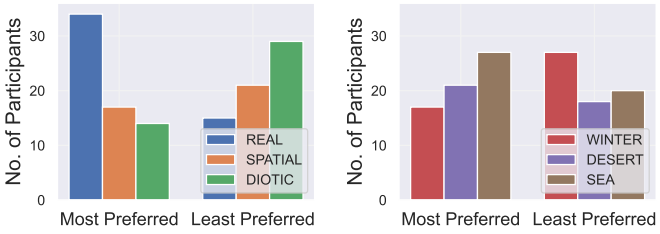


(a) "The conversation felt natural." (b) "The elem. of the env. were all of the same quality." (c) "The environment sounded convincing."

Fig. 5: Plausibility survey ratings of three exemplary items.

preferences were analyzed using subjective questionnaire-based evaluation. While the spatial audio condition showed tendencies for an increase in those metrics over the diotic condition, a significant effect of the auralization method was not found. Communication in the real world received significantly higher ratings, and was overall most preferred.

To better understand the impact of spatial audio on VR communication, we plan to analyse data collected in the study using verbal and non-verbal behavioral metrics. In follow-up studies, we aim to further increase acoustic scene and communication complexity by increasing the number of interlocutors. We will also investigate the influence of visual realism by employing volumetric avatars alongside spatial audio. Knowledge gained will support informed allocation of computational resources when designing social VR applications.



(a) Condition Ranking (b) Scenario Ranking

Fig. 6: Participant preferences from rankings by trial number.

References

[1] J. Skowronek et al. “Quality of Experience in Telemeetings and Videoconferencing: A Comprehensive Survey”. In: *IEEE Access* 10 (2022), pp. 63885–63931. DOI: 10.1109/ACCESS.2022.3176369.

[2] H. J. Smith and M. Neff. “Communication Behavior in Embodied Virtual Reality”. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. Ed. by CHI. New York, NY: ACM, 2018, pp. 1–12. DOI: 10.1145/3173574.3173863.

[3] J. Li et al. “Evaluating the User Experience of a Photorealistic Social VR Movie”. In: *Int. Symp. on Mixed and Augmented Reality (ISMAR)*. IEEE. 2021, pp. 284–293.

[4] P. Sykownik et al. “VR Almost There: Simulating Co-located Multiplayer Experiences in Social Virtual Reality”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. ACM, Apr. 2023. DOI: 10.1145/3544548.3581230.

[5] 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. New York, NY, USA: Association for Computing Machinery, 2001, pp. 166–173. DOI: 10.1145/365024.365092.

[6] Janto Skowronek and Alexander Raake. “Assessment of Cognitive Load, Speech Communication Quality and Quality of Experience for spatial and non-spatial audio conferencing calls”. In: *Speech Communication* 66 (2015), pp. 154–175. DOI: 10.1016/j.specom.2014.10.003.

[7] A. Raake et al. “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.

[8] K. Nowak et al. “Hear We Are: Spatial Audio Benefits Perceptions of Turn-Taking and Social Presence in Video Meetings”. In: *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work*. Ed. by S. Boll et al. New York, NY, USA: ACM, 2023, pp. 1–10. DOI: 10.1145/3596671.3598578.

[9] C. Hendrix and W. Barfield. “The Sense of Presence within Auditory Virtual Environments”. In: *Presence: Teleoperators and Virtual Environments* 5.3 (1996), pp. 290–301. DOI: 10.1162/pres.1996.5.3.290.

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

[11] A. N. Moraes, R. Flynn, and N. Murray. “Analysing Listener Behaviour Through Gaze Data and User Performance during a Sound Localisation Task in a VR Environment”. In: *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, Oct. 2022. DOI: 10.1109/ismar-adjunct57072.2022.00102.

[12] M. E. Latoschik and C. Wienrich. “Congruence and Plausibility, Not Presence: Pivotal Conditions for XR Experiences and Effects, a Novel Approach”. In: *Frontiers in Virtual Reality* 3 (June 2022). DOI: 10.3389/frvir.2022.694433.

[13] F. Biocca, C. Harms, and J. Gregg. “The Networked Minds Measure of Social Presence: Pilot Test of the Factor Structure and Concurrent Validity”. In: *4th annual International Workshop on Presence, Philadelphia* (Jan. 2001).

[14] F. Immohr et al. “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-T Rec. P.1301. *Subjective quality evaluation of audio and audiovisual multiparty telemeetings*. 2017.

[16] J. Skowronek. “Quality of experience of multiparty conferencing and telemeeting systems”. en. PhD thesis. 2017. DOI: 10.14279/DEPOSITONCE-5811.

[17] Tivian XI GmbH. *Unipark*. Accessed 08.04.2024. URL: www.unipark.com.

[18] Exit Games Inc. *Photon Unity Networking*. Accessed 09.04.2024. URL: https://www.photonengine.com/.

[19] 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.

[20] A. Neidhardt et al. “Flexible python tool for dynamic binaural synthesis applications”. In: *Audio Engineering Society Convention 142*. Audio Engineering Society. 2017.

[21] Unity Technologies. *Unity Documentation: Audio Spatializer SDK*. Accessed 09.04.2024. URL: https://docs.unity3d.com/Manual/AudioSpatializerSDK.

[22] C. Armstrong et al. “A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database”. In: *Applied Sciences* 8.11 (2018). DOI: 10.3390/app8112029.

[23] A. Wabnitz et al. “Room acoustics simulation for multichannel microphone arrays”. In: *Int. Symp. on Room Acoustics, ISRA, Melbourne, Australia*. 2010.

[24] 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.

[25] 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.