APlausE-MR: Investigating Multi-Party Communication in Audiovisual Mixed-Reality Environments

Felix Immohr[1], Gareth Rendle[2], Annika Neidhardt[3], Anton Lammert[2], Karlheinz Brandenburg[3], Bernd Froehlich[2], Alexander Raake[1]

[1] *Audiovisual Technology Group, Technische Universität Ilmenau, 98693 Ilmenau, Germany,*
Email: {felix.immohr, alexander.raake}@tu-ilmenau.de

[2] *Virtual Reality and Visualization, Bauhaus-Universität Weimar, 99423 Weimar, Germany,*
Email: {gareth.rendle, anton.benjamin.lammert, bernd.froehlich}@uni-weimar.de

[3] *Electronic Media Technology Group, Technische Universität Ilmenau, 98693 Ilmenau, Germany,*
Email: {annika.neidhardt, karlheinz.brandenburg}@tu-ilmenau.de

Introduction

Mixed Reality (MR) technology allows people to meet and communicate in shared interactive virtual environments (IVEs), providing opportunities for rich social interaction between remote parties, as well as enabling researchers to study inter-personal communication in laboratory scenarios that are controlled, yet highly realistic. Creators often strive to develop plausible IVEs that match our expectations of real-world communication, by supporting high-fidelity audiovisual user representations and virtual environments. The **A**udiovisual **Plaus**ibility and **E**xperience in Multi-Party **M**ixed **R**eality project (APlausE-MR), part of DFG SPP2236 AUDICTIVE, aims to investigate how characteristics of audiovisual IVEs contribute to user experience in MR. Note that we follow [21] by defining MR to include both Virtual Reality (VR) and Augmented Reality (AR). We hypothesize that, in communication contexts, audiovisual factors can work together in a mutually reinforcing way, such that imperfections in one modality can be compensated for by other factors. The project's research challenges address identification of factors affecting experiences in multi-party IVEs, as well as the development of technologies and evaluation methods required to quantify the influence of those factors.

This document gives an overview of the research and development contributions of the APlausE-MR project to date, which can be summarized as:

- development of a VR study execution framework that allows remote participants to communicate in a realistic audiovisual IVE;
- a collaborative immersive analytics studio and conversational analysis framework that allow recording and post-hoc analysis, re-exploration and annotation of VR experiments;
- technical advancements that allow volumetric avatar reconstruction at a higher frame rate than the capture frame rate of cameras used to capture the subject;
- two studies examining the effect of spatial audio on users' experience in multi-party VR communication scenarios, including the initialization of a questionnaire to measure plausibility.

Study Execution and Analysis Framework

Research in the APlausE-MR project is centered around multi-party communication studies in IVEs. A study framework was developed that enables multiple participants to meet and communicate in a shared virtual environment. Software was developed to record the virtual environment and events that occurred during each study run for subsequent analysis.

Study IVE

The IVE developed for APlausE-MR studies allows participants at remote locations to meet in a shared virtual environment. Communication is supported by transmission of voice signals and participant's movements, which drive virtual user representations (i.e. avatars) that appear in the virtual environment. Users may appear as abstract avatars, with simple head, torso and hand models that are animated, based on the tracked positions of a head-mounted display and associated controllers; or as live-captured, photo-realistic volumetric avatars. The framework supports the use of realistic virtual environments, such as those reconstructed using photogrammetric methods.

To identify the influence of different audio spatialization approaches, a system that allows for fine-grained control over the audio spatialization pipeline was implemented. Realistic spatial audio is provided by interfacing with an extended version of the *pyBinSim* binaural simulation library [15]. Spatialized audio sources in the Unity scene send audio data via a Native Audio Plugin[1] to a pyBinSim server that applies binaural simulation processing. PyBinSim synthesizes direct sound dynamically, based on the SADIE II HRTF data set (subject D2 - Kemar) [1] and the mouth directivity dataset *female speech* provided with the toolbox MCRoomSim [26]. The reverberation, based on a BRIR measured with a KEMAR 45ba head-and-torso simulator, is not position-dynamic, since it was shown that within a certain area in front of the sound source this approach leads to an impression as

[1] docs.unity3d.com/Manual/AudioSpatializerSDK.html

plausible as an entirely measured BRIR dataset [16].

The IVE is implemented in the Unity game engine, using Photon Voice for voice transmission, and Photon Unity Networking for distribution of the scene state[2]. The experiment flow is controlled using the bmlTUX framework [3].

Immersive Analytics Studio

Analysis of users' verbal and non-verbal behavior in IVEs can be used to quantify the influence of environment factors [24, 7]. For this reason, detailed recordings of user behavior are valuable, supporting in-depth analysis. Various systems that facilitate recording user behavior in IVEs and immersive analysis through playback of recorded data have been proposed [9, 8, 12]. However, existing systems are limited in that they do not support collaborative analysis, which is a key benefit of immersive analytics systems [6]. For the APlausE-MR project, we developed the *Collaborative Immersive Analytics Studio* (CIAS), which enables the collaborative analysis of user behavior recordings. Synchronous recording of participant's microphone and speaker audio, as well as of changes in their respective virtual scenes, can be created. Recorded data can be read and replayed to facilitate immersive collaborative re-exploration and analysis of user behavior in the virtual environment. Immersive re-exploration means that researchers can experience playback of recordings in immersive VR while being able to navigate freely in the IVE.

The CIAS provides tools for both automated and manual analysis of recorded user studies. To support manual identification and classification of certain behaviors (behavioral coding), analysts have the ability to create and share annotations for temporal intervals that can be associated with particular objects in the virtual scene. Temporal navigation to the start of a selected annotation is supported to reduce the need to search for temporal points of interest.

Automated analysis of study data is provided through a query interface. Analysts can create queries using a query language similar to that proposed by Marquardt et al. [14], resulting in automated detection of events based on distance, velocity, containment, gaze, audio level, and movement criteria. Detected events are visualized as annotations that are shared between analysts.

As the playback of recorded spatialized audio in a virtual environment does not necessarily provide a correct spatial audio experience to analysts who can navigate freely in the virtual environment, re-spatialization of recorded user audio is required to provide analysts with an audio experience that is as close as possible to the original study context. This is implemented by recording the microphone audio of each study participant, which is subsequently aligned for synchronous playback and spatialized during analysis sessions.

[2] photonengine.com

Volumetric Avatar Research

Volumetric avatars are photo-realistic user representations that are reconstructed and transmitted to remote locations in real time [5, 25]. Their ability to capture full-body movements and gestures facilitates rich communication in social MR scenarios. The APlausE-MR project aims for development of an IVE that includes highly realistic user representations. To this end, we propose an approach for doubling the reconstruction frame rates of volumetric avatars, using commercially available RGBD cameras [19]. Increasing the reconstruction frame rate, in our case from 30 to 60 FPS, aims to achieve smoother reconstructed motion, and avoid judder effects when avatars are rendered on VR displays with refresh rates of 60+ FPS. Our method divides the available cameras into two capture groups. While all cameras capture RGBD images at 30 FPS, one group is temporally offset, such that a 60 FPS stream is created. To avoid flickering artifacts, temporal fusion is applied to produce a temporally coherent volumetric avatar stream.

Multi-Party Virtual Reality Communication Studies

Due to the complex multi-modal nature of IVEs, there are many system characteristics that can affect users' perception and experience. These include technical factors, like visual and auditory fidelity, and the system's support for natural interactions [23]. To evaluate and compare immersive systems, researchers have developed generalizable constructs. One such construct, *social presence*, is commonly used in the context of communication systems, and has also been investigated in the context of IVEs [24, 13]. While traditional video conferencing systems and associated evaluation methods have been thoroughly studied and standardized, factors influencing audiovisual plausibility and social presence in multi-modal IVEs need to be investigated further. To this end, suitable evaluation methods need to be identified and established.

One factor shown to positively impact experience in communication scenarios is spatial auditory representation, both in traditional conferencing systems [2, 23, 18, 22] for listening-only and conversation tests as well as more immersive systems, although only in passive listening-only studies [17]. In APlausE-MR, we began by investigating the effect of realistic spatial audio as one influencing factor in an interactive conversational context by conducting two interactive communication studies in multimodal VR.

Study 1: Dyadic negotiation scenario

We conducted an initial proof-of-concept study to evaluate the influence of spatial audio on users' experience during an interactive two-party conversation in VR. Participants completed a within-subject negotiation task paradigm adapted from [24], based on a visual reference, in three conditions: a real-world face-to-face condition (F2F); a VR condition with spatial audio, realized through position-dynamic binaural synthesis; and a VR

condition with purely diotic audio representation. The study setup was based on the Study Execution Framework as described above. For evaluation, social presence and participant preferences were assessed subjectively through questionnaires and an informal post-study interview. For objective analysis, behavioral assessment was performed based on speech and camera video recordings. This included the assessment of non-verbal behavior through gestural analysis and the conversational turn-taking behavior based on a parametric conversation analysis approach.

This initial study, which served as proof-of-concept for the study execution framework, revealed tendencies for improvements with binaural over diotic audio regarding social presence, without significant effect on the indirect measures, potentially due to the chosen communication scenario. F2F interaction could be distinguished by subjective and objective metrics. While most participants indicated that F2F was most preferred, the VR condition without spatial audio was least preferred. A contribution detailing this work was presented at the ACM IMX 2023 conference [10].

Study 2: Dyadic collaboration scenario

Building on the insights of the first study, a second study investigates the effect of binaural spatial audio on plausibility, social presence and communication behavior in IVEs. We revised the test method by adjusting the task and the subjective measurement instrument in the form of an adapted post-trial questionnaire. Therefore, the study aimed to provide insight into the suitability of this test method to discriminate conditions within the same communication medium. The test employed a 'spot-the-differences' task, which is a VR adaptation of the Leavitt task used in traditional video conferencing assessments [11]. Participants were instructed to identify differences between colored shapes represented individually on boxes placed in the virtual environment. The boxes were distributed to encourage participant movement and emphasize the use of spatial hearing. The test consisted of four conditions, comparing spatial to diotic audio in two different scenes. To extend upon existing, often context-dependent, methods for evaluation, a questionnaire was designed to enable the assessment of quality and plausibility aspects. This included sub-dimensions related to task, enjoyment, interaction, quality and coherence as derived from the literature. Participants completed the questionnaire, as well as an established social presence inventory, after each trial. In addition to subjective ratings, the whole scene, including participant behavior, was recorded using the CIAS as detailed above. After the final condition, a short active listening preference test was conducted in VR, where one participant was prompted to speak, and the second participant was instructed to actively listen and switch between the audio conditions, before providing a preference indication based on not associable labels.

The study revealed that the holistic subjective metrics of social presence and plausibility used could not distinguish the conditions in a conversational situation. Therefore, a significant impact of spatial audio over a diotic representation on these metrics could not be shown. However, in the direct comparison used as second part of each test run, spatial audio was significantly preferred. A publication of this study, including in-depth conversation and behavioral analysis, is currently under preparation.

Conclusion and Future Work

In the APlausE-MR project, we have developed a high-fidelity audiovisual IVE for conducting, recording, and analyzing studies. The studies performed so far have given us insight into evaluation methods for multi-party communication scenarios in VR, and are set to inform the design of subsequent studies.

Results from the described studies suggest that spatial audio has little effect on holistic measures of the users' experience, like responses to social presence questionnaires and objective metrics derived from behavioral analysis. One potential explanation for this is the relative simplicity of the communication scenario presented to participants. Both study scenarios featured one-on-one communication, limiting the auditory scene complexity. In upcoming studies, additional participants will be included, emphasizing the use of spatial hearing for discerning which conversation partner is speaking, assuming a greater effect on user experience in scenarios with more than two speakers.

The work planned for the remainder of the project focuses on further studies that expand the existing scenarios by including higher communication complexity, e.g., with more participants, extending the variety of MR displays used, and including realistic user representations in the form of volumetric avatars.

APlausE-MR aims to examine the effect of visual VE characteristics on users' experience, as well as the effect of audio VE characteristics. Accordingly, subsequent studies will integrate volumetric avatars into the study scenarios. This will enable comparison of volumetric avatars against abstract avatars, as well as supporting closer investigation of the importance of volumetric avatar characteristics such as frame rate and reconstruction resolution.

Study scenarios will also be expanded to include a second type of MR visual display, namely stereoscopic projection screens. The projection screens allow true mixed-reality scenarios, where one collocated group, viewing the virtual scene through one screen, can communicate with a remote group viewing the virtual scene through a second projection screen in a separate physical location [4]. This configuration gives us the opportunity to study how communication is affected when some communication partners are remote, while others are located in a shared physical space.

References

[1] C. Armstrong et al. (2018): A perceptual evaluation of individual and non-individual hrtfs: a case study of the sadie ii

database. In: Applied Sciences, 8, 11.

[2] J. Baldis (2001) Effects of spatial audio on memory, comprehension, and preference during desktop conferences. SIGCHI conference on Human factors in computing systems, S. 166-173.

[3] A. O. Bebko and N. F. Troje (2020) bmlTUX: Design and Control of Experiments in Virtual Reality and Beyond. I-Perception, 11(4).

[4] S. Beck et al. (2013): Immersive Group-to-Group Telepresence. In: IEEE Transactions on Visualization and Computer Graphics, 19(4).

[5] M. Dou et al. (2016) Fusion4D: Real-time performance capture of challenging scenes. ACM Transactions on Graphics, 35(4).

[6] B. Ens et al. (2021): Grand Challenges in Immersive Analytics. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. S. 1-17.

[7] G. Gorisse et al. (2019): From Robot to Virtual Doppelganger: Impact of Visual Fidelity of Avatars Controlled in Third-Person Perspective on Embodiment and Behavior in Immersive Virtual Environments. In: Frontiers in robotics and AI 6, S. 8.

[8] S. Howie and M. Gilardi (2021): Virtual Observations: a software tool for contextual observation and assessment of user's actions in virtual reality. In: Virtual Reality 25 (2), S. 447-460.

[9] S. Hubenschmid et al. (2022): ReLive: Bridging In-Situ and Ex-Situ Visual Analytics for Analyzing Mixed Reality User Studies. In: CHI Conference on Human Factors in Computing Systems. S. 1-20.

[10] F. Immohr et al. (2023): Proof-of-Concept Study to Evaluate the Impact of Spatial Audio on Social Presence and User Behavior in Multi-Modal VR Communication. Accepted to ACM IMX 2023.

[11] ITU-T Rec. P.1301 (2017): Subjective quality evaluation of audio and audiovisual multiparty telemeetings. International Telecommunication Union.

[12] S. Kloiber et al. (2022): VR-based Exploration of Participant Movement in Experimental Psychology. In: 2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). S. 391-396.

[13] J. Li et al. (2021): Evaluating the user Experience of a Photorealistic Social VR Movie. In: 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). S. 284-293.

[14] N. Marquardt et al. (2015): EXCITE: EXploring Collaborative Interaction in Tracked Environments. In: Human-Computer Interaction - INTERACT 2015, Bd. 9297. Cham: Springer International Publishing (Lecture Notes in Computer Science), S. 89-97.

[15] A. Neidhardt et al. (2017): Flexible Python Tool for Dynamic Binaural Synthesis Applications. In: Engineering Brief 346.

[16] A. Neidhardt et al. (2018): Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets. In: 144h Int. AES Convention, Milan, Italy.

[17] T. Potter et al. (2022): On the Relative Importance of Visual and Spatial Audio Rendering on VR Immersion. In: Frontiers in Signal Process. Audio and Acoustic Signal Processing.

[18] A. Raake et al. (2010): Listening and Conversational Quality of Spatial Audio Conferencing. In: Audio Engineering Society Conference: 40th Int. Conf.: Spatial Audio: Sense the Sound of Space.

[19] G. Rendle et al. (2023): Volumetric Avatar Reconstruction with Spatio-Temporally Offset RGBD Cameras. Presented at the 2023 IEEE Conference on Virtual Reality and 3D User Interfaces.

[20] H. Sacks et al. (1974): A simplest systematics for the organization of turn-taking in conversation. In: Language, 50:696-735.

[21] R. Skarbez et al. (2021): Revisiting Milgram and Kishino's Reality-Virtuality Continuum. In: Frontiers in Virtual Reality 2. S. 2673-4192.

[22] J. Skowronek and A. Raake (2015): Assessment of Cognitive Load, Speech Communication Quality and Quality of Experience for spatial and non-spatial audio conferencing calls. In: Speech Communication, 66, S. 154-175.

[23] J. Skowronek et al. (2022): Quality of Experience in Telemeetings and Videoconferencing: A Comprehensive Survey. In: IEEE Access.

[24] H. J. Smith and M. Neff (2018): Communication Behavior in Embodied Virtual Reality. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18), 1-12.

[25] D. Tang et al. (2020). Deep Implicit Volume Compression. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1290-1300.

[26] A. Wabnitz et al. (2010): Room acoustics simulation for multichannel microphone arrays. In: Int. Symp. on Room Acoustics, ISRA, Melbourne, Australia.