Influence of Audiovisual Realism on Communication Behaviour in Group-to-Group Telepresence

Gareth Rendle [*]
Virtual Reality and Visualization,
Bauhaus-Universität Weimar, Germany

Felix Immohr [†]
Audiovisual Technology Group, Technische Universität Ilmenau, Germany

Christian Kehling [‡]
Electronic Media Technology Group, Technische Universität Ilmenau, Germany

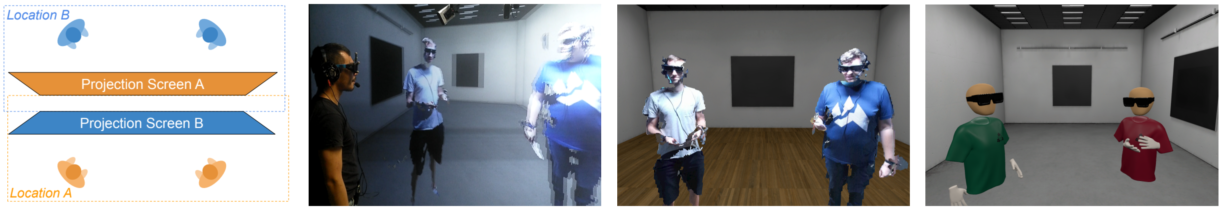
Anton Lammert [§]
Virtual Reality and Visualization,
Bauhaus-Universität Weimar, Germany

Adrian Kreskowski [¶]
Virtual Reality and Visualization,
Bauhaus-Universität Weimar, Germany

Karlheinz Brandenburg [‖]
Electronic Media Technology Group, Technische Universität Ilmenau, Germany

Alexander Raake [**]
Audiovisual Technology Group, Technische Universität Ilmenau, Germany

Bernd Froehlich [††]
Virtual Reality and Visualization,
Bauhaus-Universität Weimar, Germany



(a) Four participants in two locations. (b) Participant POV. (c) Volumetric (VOLU) avatars. (d) Abstract (ABST) avatars.

Figure 1: Groups of four participants (split into collocated pairs) communicate through a group-to-group telepresence system based on two multi-user projection screens (1a). Each person can see and hear the three other participants; one collocated, and two remote (1b). Remote participants are represented either as volumetric (VOLU) avatars (1c) or abstract ABST avatars (1d).

Abstract

Group-to-group telepresence systems immerse geographically separated groups in a shared interaction space where remote users are represented as avatars. Notably, such systems allow users to interact with collocated and remote interlocutors simultaneously. In this context, where virtual user representations can be directly compared with real users, we investigate how visual realism (avatar type) and aural realism (presence of spatial audio) affect communication. Furthermore, we examine how communication differs between collocated and remote pairs of interlocutors. In our user study, groups of four participants perform a collaborative conversation task under the aforementioned visual and aural realism conditions. Our results indicate that avatar realism has positive effects on subjective ratings of perceived message understanding and group cohesion, and yields behavioural differences that indicate more interactivity and engagement. Few significant effects of aural realism were observed. Comparisons between collocated and remote communication found that collocated communication was perceived as more effective, but that more visual attention was paid to both remote participants than the collocated user.

Index Terms: Telepresence, Mixed reality, Mediated communication, Avatars, Spatial audio.

1 Introduction

A profound shift in societal attitudes towards remote working and collaboration has increased the demand for effective communication tools for physically distributed parties. Accordingly, *telepresence* technologies have received increased interest [64, 15, 70, 27]. Telepresence systems establish a shared sense of presence among geographically separated users [11] by immersing them in a common interaction space, where remote users are embodied by avatars. The non-verbal cues that avatars can convey and the spatial context that allows understanding of those cues enable natural communication, which has proven effective in contexts where building trust, sharing emotions, and explaining ideas are crucial [70].

Remote communication often encompasses collaboration between multiple remote groups, where each group consists of two or more collocated persons. However, despite the availability of immersive multi-user telepresence technologies [34, 5, 75], few studies have been conducted in telepresence scenarios with more than one person in each collocated party [5, 50, 51], leaving many questions about group-to-group telepresence unanswered. For example, it is not known if the combination of real and virtual stimuli affects the perceived quality or plausibility of virtual stimuli, or how different modalities and characteristics of the virtual stimuli contribute to supporting communication. There are also limited guidelines for assessing the effectiveness of these communication systems.

To address this knowledge gap, we designed and conducted a

[*] e-mail: gareth.rendle@uni-weimar.de
[†] e-mail: felix.immohr@tu-ilmenau.de
[‡] e-mail: christian.kehling@tu-ilmenau.de
[§] e-mail: anton.benjamin.lammert@uni-weimar.de
[¶] e-mail: adrian.kreskowski@uni-weimar.de
[‖] e-mail: karlheinz.brandenburg@tu-ilmenau.de
[**] e-mail: alexander.raake@tu-ilmenau.de
[††] e-mail: bernd.froehlich@uni-weimar.de

group-to-group telepresence user study. Four participants, two in each collocated group, performed a conversation task while using a telepresence system based on two multi-user projection screens located at different sites. The visual and aural realism of the remote users' representations were controlled as an independent variable, with participants performing a collaborative activity that simulates a goal-oriented meeting under each condition. We collected participants' impressions of communication with collocated and remote colleagues, and recorded movement and speech data for analysis.

With the subjective and objective data from our user study, we aim to answer three research questions (RQs). **RQ1** is: **how does the visual realism of remote users' avatars affect communication in group-to-group telepresence?** Research has indicated that *volumetric* avatars, which are generated in real-time from colour and depth image streams, and therefore exhibit high levels of realism, are preferred over avatars created by animating computer-generated 3D models [73, 13, 16, 3, 74, 33]. However, this has not been verified in a group-to-group telepresence context, where virtual stimuli can be compared directly against real users.

RQ2 is: **how does spatial processing of remote users' voices affect communication in group-to-group telepresence?** Spatial audio has shown positive effects in audio-only and video-conferencing communication and non-interactive immersive contexts [4, 55, 32, 63, 48], but the methods and scenarios traditionally used to quantify its positive influence have not been able to elicit measurable effects in immersive communication systems [24, 25, 26]. The group-to-group telepresence context presents a scenario with multiple audio sources, and where virtual sources can be directly compared to real sources, potentially increasing the measurable impact of spatial audio.

Finally, **RQ3** is: **how does communication between collocated users differ from communication between remote users in group-to-group telepresence?** In an ideal system, the medium should disappear; that is, a user's communication with remote colleagues should not differ from that with collocated colleagues. Identifying differences and their underlying causes would support the development of effective telepresence systems.

This work provides the following contributions:

- The design, implementation and evaluation of a group-to-group telepresence user study, in which groups of four participants, distributed in pairs between two remote physical locations, communicate to solve a collaborative conversation task under different virtual environment conditions.
- An investigation of the effect of two facets of the virtual user representation (avatar appearance and spatial audio) on communication behaviour and user experience, finding evidence that employing volumetric avatars instead of abstract tracked avatars improves perceived message understanding and group cohesion, and yields behavioural differences that indicate more interactive communication.
- A comparison of communication behaviour between remote and collocated pairs of participants, finding that collocated communication was perceived as more effective, but that more visual attention was paid to remote participants.

Our work informs the development of effective group-to-group telepresence systems and provides a reference for future analysis of simultaneous communication with collocated and remote users.

2 Related Work

While the meaning of the term *telepresence* has evolved over time, the definition most relevant in this work is "the use of technology to establish a sense of shared presence or shared space among geographically separated members of a group" [11]. The sense of shared presence is usually fostered by representing remote users virtually as avatars in a shared interaction space. Immersion in the shared environment means that users can contextualise directional non-verbal cues conveyed by the avatars, like pointing gestures and eye movements. Telepresence technologies for remote communication between individuals have received attention for decades [34, 19, 42, 53, 44] and have attained impressive realism by reconstructing high-fidelity, real-time user representations [49, 40, 67].

2.1 Group-to-Group Telepresence

Most telepresence systems only support one user per physical location. In contrast, group-to-group telepresence systems allow groups of collocated users to communicate with other groups in remote locations. A prerequisite is that each collocated user receives their own perspective-correct view of the virtual stimuli. This, along with the challenge of reconstructing and transmitting multiple user representations, means that group-to-group virtual communication has received less attention than one-to-one systems.

Beck et al. developed a group-to-group telepresence system that allows two groups of up to six collocated users to meet in a shared virtual workspace [5]. Each user sees their own perspective-correct stereoscopic view of the virtual scene on a multi-user projection screen [38]. More recently, consumer Head-Mounted Displays (HMDs) that enable the real environment and spatially anchored virtual content to be perceived simultaneously have been released[1][2]. By registering many devices into a shared coordinate system, collocated users perceive consistent views of the virtual scene. Irlitti et al. [27] perform a qualitative analysis of asymmetric HMD telepresence scenario, finding that collocated and remote collaborations were based on similar communication techniques.

Unlike previous work on group-to-group telepresence, we investigate the effect of visual and aural realism, while extending qualitative analyses of communication in existing works by extracting quantitative behavioural and conversational metrics.

2.2 Avatar Realism in Telepresence

The effect of avatar realism has received attention in previous work on one-to-one telepresence and virtual humans more generally, where a distinction between *behavioural* realism (how much the avatar moves like their physical counterpart) and *appearance* realism (how much the avatar looks like a human) is drawn [8]. Avatars can be categorised into *tracked* avatars, which are pre-generated models that are animated by the user's movements, and *reconstructed* avatars, which are generated from real-time image streams [65]. The increased accessibility of technologies for real-time reconstruction of avatars, including the availability of consumer depth cameras, means that a high degree of visual and behavioural realism is now possible for reconstructed avatars, also known as *volumetric* avatars [13, 12, 56]. The 3D models required for tracked avatars can also be created quickly using photogrammetry techniques and animated by data from a combination of sensors that track body movements, facial expressions and gaze direction [74].

Various works have compared volumetric avatars to tracked avatars, finding that volumetric avatars perform better in terms of social presence [74, 13, 33], collaborative task performance [16] (although this may be task-dependent [74]), trustworthiness [3] and positive affinity [73]. Tracked avatars often suffer from limited behavioural realism, due to inadequate reconstruction of facial expressions and eye movements [3, 13, 16, 33, 73]. When those features were present [74], participants still rated volumetric avatars higher on humanness and social presence, perhaps due to the remaining deficiencies in tracking quality. However, volumetric avatars incur additional computational and hardware costs, meaning that it is important to quantify their effect in a group-to-group setting, where reconstruction of multiple users magnifies those costs.

[1] https://www.microsoft.com/hololens
[2] https://www.meta.com/quest

2.3 Spatial Audio

Spatial audio leverages interaural differences to create the impression that sounds are produced by a source that is placed in a particular location. In audio- and video-conferencing systems, spatial audio has been shown to have positive effects on memory and comprehension [4], intelligibility and listening effort / cognitive load [55, 32], and interactivity and social presence [48]. In immersive telepresence systems, spatial processing can be applied to the voices of remote users to localize them in the shared space. Studies in immersive systems have shown positive effects of spatial audio on presence [22], social presence [57], and psychological immersion [54], although these results are from perception-only studies where participants do not communicate during the test. Previous work that did investigate spatial audio in immersive communication contexts did not uncover significant effects on participant behaviour or reported social presence, either in dyadic [26, 25] or triadic [24] conversational contexts. In this work, we examine the effect of spatial audio in a group-to-group telepresence context where multiple audio sources, real and virtual, are present in a reverberant room. We investigate whether the higher degree of communication and acoustic scene complexity leads to a greater utility of spatial audio, therefore producing a measurable effect on communication-related participant ratings and behavioural metrics.

2.4 Behavioural Analysis in Telepresence

Quantifying users' verbal and non-verbal actions produces objective measures of communication behaviour that supplement subjective questionnaire responses (which may be affected by the cognitive load induced by a communication task). Links between different behaviours and high-level user states [59] and constructs like social presence [21] have been explored. Indicators that can be extracted from recorded tracking and audio data without manual intervention include proxemics states [71, 20], movement synchrony [43], conversational structure metrics [10, 61, 55, 60], and gaze behaviour. Faster conversational turn-taking, a greater degree of overlap between speakers, and shorter pauses between turns are associated with increased interactivity [61]. Gaze direction can be used to extract metrics quantifying how much time is spent looking at different gaze targets, and how often focus is shifted between gaze targets. Amount of fixation on a target has been used to quantify the degree of engagement during social interactions [28], while shorter fixations have also been associated with engagement, as people scan interlocutors' faces [6]. Otsuka [51] performs a thorough behavioural analysis of recorded conversations in a communication scenario where remote participants are embodied by kinetic robots, comparing behaviour between mediated and unmediated contexts. We take a similar approach in using behavioural metrics to compare communication between experimental conditions, and to compare collocated and remote interactions.

3 User Study

To study communication in group-to-group telepresence, we designed a four-party user study in which each participant interacted with collocated and remote users simultaneously. The participants in each group were split into two collocated pairs.

The group-to-group scenario requires that participants perceive perspective-correct virtual stimuli (representing the remote users) and the real surroundings (including the collocated user) simultaneously. This is made possible in this work by a multi-user stereoscopic projection system based on Digital Projection INSIGHT 4k HFR 360 projectors. A key benefit of this system is that real stimuli can be seen directly through the active shutter glasses, in contrast with video see-through head-worn Mixed Reality (MR) devices that capture and reproduce the real environment on their display, potentially suffering from image warping. Optical see-through MR devices could be used in a similar way, but their field-of-view is typically quite limited. Allowing unmediated perception of the real environment means that we can evaluate the effect of combining real and virtual stimuli in a communication scenario. More details about the technical setup are given in Section 3.5.

To simulate a meeting scenario, participants solve a series of collaborative discussion tasks, explained in more detail below. The realism of the visual and aural representations of remote users was varied systematically in each trial of the study. While participants engaged in the task, their movements and voice signals were recorded. After each discussion task, participants responded to questionnaires. The following sections describe the study design and technical setup in more detail. Approval for the study was obtained from the ethics committee of TU Ilmenau (*Ethikkomission der TU Ilmenau*).

3.1 Independent Variables

In line with our research questions, the two independent variables controlled during the study were the visual realism of the remote users' avatars and the aural spatial realism of their voice signals.

Conducting a joint investigation into the effect of visual and aural realism factors means that results can reveal the relative impact of changes to each variable, as well as potential audiovisual interaction effects, which have been observed in video quality assessments [17]. This information can guide allocation of computation resources when developing MR applications for communication.

The avatar realism variable had two levels, shown in Figures 1c and 1d. Remote users were either represented as volumetric avatars, reconstructed in real time using images from colour and depth cameras (labelled as the VOLU condition), or less realistic, "abstract" avatars, consisting of simple head, hand and torso geometry that was animated by the motions of the user (ABST). These levels provide two strongly contrasting levels of realism. The volumetric avatars have a higher degree of appearance and behavioural realism, in that they appear more like a human (all body parts are present, and are coloured correctly with respect to the real person) and convey most of the users' movements. We note that the eyes of the volumetric avatars are occluded due to the shutter glasses that must be worn to use the projection system. Since the abstract avatars are animated by the users' movements, they can convey head and hand movements, but not gaze direction, facial expressions, or movements of other parts of the body (including finger movements). The appearance and behavioural realism of the abstract avatar are both low to avoid uncanny impressions that may arise when appearance and behavioural realism levels are not coherent. By choosing avatar types with contrasting levels of realism, we aimed to elicit noticeable differences in subjective and objective measures.

The two types of aural spatial realism are: the spatialised condition (SPAT), where binaural processing was applied to remote users' voices, supporting localisation of the sound in the respective real environment; and the diotic condition (DIOT), where no spatial processing is applied, with listeners receiving the same signal in each ear. Like the visual realism variable, we chose strongly contrasting levels of aural spatial realism. During the study, each group performed one trial for each of the four combinations of conditions (VOLU+SPAT, VOLU+DIOT, ABST+SPAT, ABST+DIOT).

3.2 Conversation Task

During each trial, participants completed a version of the Survival Task from ITU-T Rec. P.1301 Appx.VI [30, 62], which was originally designed for the assessment of traditional telemeeting systems. Before the trial, each participant was given the description of a 'survival' scenario and a list of three items. The group then had the task to select six out of the twelve objects that would best help them survive in the given scenario. Scenario descriptions and items for the *mountains*, *sea*, *moon*, *desert*, and *winter* scenarios were used from the recommendation. Some item descriptions and images were adjusted for clarity. The scenario descriptions and item lists used during the experiment are shown in Appendix B.

Participants were given one minute before each discussion task to read the scenario and their item list, provided on an A6 piece of paper. The group was given 8 minutes to select their items, with an audible warning signal played after 7 minutes. Participants could hold the scenario description and item paper throughout the trial. The study supervisor monitored the conversation and ended the trial if a consensus was reached between the participants.

The survival task was chosen because it instigates more natural, free-flowing discussion between group members than structured conversational tasks where turn-taking is more rigid (such as the celebrity name guessing game) [30]. This made the task a good approximation of a work meeting where colleagues collaborate to decide on a strategy by exchanging views and finding a consensus. The distribution of items between participants meant that each group member had to contribute to the discussion, mitigating the dominance of more talkative group members by encouraging speaking time to be split more evenly among group members. Participants were given time to read the task before the trial to avoid silent reading time during the trials. The purely conversational nature of the task meant that movement was not required, which was beneficial due to the limited extent of the avatar capture space.

3.3 Data Collection

Throughout each trial, audio from participants' microphones was recorded, along with the tracked poses of their head and hands, for behavioural and conversational analysis.

After each trial, each participant answered a questionnaire made up of three sections: part one, addressing communication with their collocated participant; part two, addressing communication with the remote users; and part three, which addressed general group communication during the trial. In parts one and two, participants indicated their level of agreement with statements relating to Perceived Message Understanding, a subscale of the Networked Minds questionnaire [7]. Part two also contained six additional statements about the remote users' visual and audio representations. In part three, six statements queried the perceived naturalness, sense of togetherness, and success of the conversation with respect to the conversation task. Additionally, participants rated the overall experience during each trial and indicated their current level of discomfort, yielding 26 questions in total (shown in Table 1). A 7-point Likert scale was used for all questions, where 7 indicated that they agreed with the given statement. Question order was fixed throughout the experiment to maintain a clear relationship between section and participant group, at the expense of potential order effects.

After the fourth trial, participants completed a final questionnaire, in which they ranked the trials in order of preference and reported their level of familiarity with the other participants (whose identity was not known to them before the study). They could also answer an open question about their experience during the study.

3.4 Study Procedure

On arrival, participants were given an information sheet and asked to complete a consent form and a short demographic survey. After putting on the necessary equipment, participants completed an introduction round, in which they introduced themselves to the rest of the group. While introduction rounds in group conversation tests are often performed in a face-to-face context, the distance between the city locations of the projection screens made this impractical for our study. The introduction and training rounds were instead performed under the most realistic test condition (VOLU+SPAT).

Participants then completed a training round of the conversation task, followed by four trials, each with a different condition. The condition order over the four trials was varied using a balanced Latin square design. After the fourth trial, the final questionnaire was administered. Question responses were given with the UNIPARK [66] platform, using PCs that were placed apart to avoid participant communication between trials. The user study took up to 90 minutes and participation was compensated with 20€.

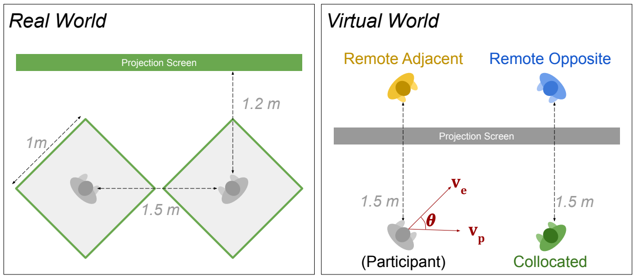


Figure 2: Participants in each location are instructed to stay within marked 1 $m \times 1$ m squares in front of the projection screen (left). The coordinate spaces of locations A and B are aligned in the virtual world so that participants are 1.5 m from their nearest remote participant (right). Each participant has one *collocated* teammate, one *remote opposite* teammate, and one *remote adjacent* teammate. During gaze analysis, the angle θ between the participant's viewing vector $\mathbf{v_e}$ and the vector from their eye to another participant's head $\mathbf{v_p}$ is used to derive gaze state.

During each trial, participants were directed to stay within marked 1 $m \times 1$ m squares. The centres of the squares in each location were 1.5m apart, and the placement of each projection screen in the shared virtual space meant that the participants' locations formed a 1.5 $m \times 1.5$ m square (see Figure 2). We henceforth refer to the person in a user's real space as their *collocated* teammate, while the others are referred to as the *remote opposite* and *remote adjacent* teammates, as shown in the image. The distance between each participant and their collocated teammate was the same as the virtual distance between the participants and their remote adjacent teammate. This arrangement aimed to prevent proximity differences from affecting the interaction between participants. Participants occupied the same square for all trials, so always had the same collocated partner, because the distance between the lab sites meant that changing locations during the study was impractical.

3.4.1 Participants

A total of 18 groups completed the study. Two study runs were excluded due to technical problems interrupting the experiment, leaving 16 groups made up of 64 participants (33 male, 29 female, 2 diverse) aged 23-36 years (M=26.9, SD=3.1). 14 groups were mixed in gender, one was all-male and one all-female. The level of familiarity between participants was generally low, with 80.2 % of participants' responses indicating "no familiarity" when asked to rate familiarity with teammates. However, there were only 3 groups where participants reported no familiarity with each other at all.

3.5 Study Setup

The user study was based on a Unity application built with Unity version 2021.3.26f1. One PC per participant hosted an instance of the application, while a further instance ran on a control PC, enabling the study supervisor to control the experiment flow. PCs were equipped with an Intel(R) Core(TM) i9-9900X CPU at 3.50 GHz, and a NVIDIA Quadro RTX 6000 GPU. Distribution logic was handled by the *VRSYS-Core* framework[3], which in turn leverages Unity Netcode to distribute game object transforms (i.e.

[3] https://github.com/vrsys/vrsys-core

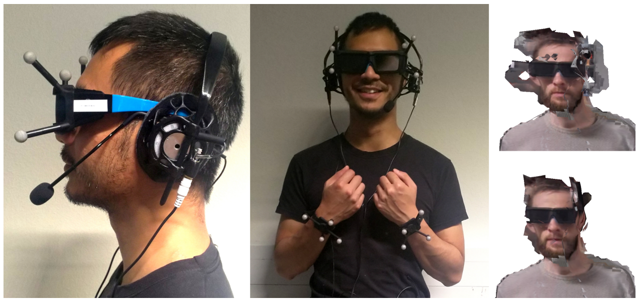


Figure 3: Equipment worn by each user: tracked shutter glasses, microphone, headphones, and hand-tracking targets. Right: tracking marker artefacts before (above) and after mitigation (below).

the poses of the abstract avatars' head, hand and torso geometry) and ODIN 4Player voice chat SDK to transmit participants' voices[4]. An open-source recording plugin was used to record participants' head and hand poses and microphone signals [39].

In each location, two participants stood in front of a multi-user stereoscopic projection screen (of size $4.32\,m \times 2.32\,m$ in location A and $5.00\,m \times 2.65\,m$ in location B), on which their remote conversation partners' avatars were displayed. The projection system provides each user with individual, perspective-correct stereo image streams at 60 Hz. Participants wore Volfoni Edge shutter glasses to enable separate images to be perceived by each eye. Each pair of glasses was tracked by an outside-in Infrared (IR) optical tracking system [1], required for correct projection, spatial audio rendering and animation of abstract avatars. Motion-to-photon latency of the display system was measured at 140 ms. Participants wore ART hand-tracking targets to animate the hands of their avatar, as well as microphones and headphones for audio transmission (Figure 3).

3.5.1 Avatar Reconstruction

Under the VOLU condition, avatars were reconstructed in real-time from colour and depth image streams captured at a rate of 30 Hz with respective pixel resolutions of 1280x720 and 640x576 using four *Microsoft Azure Kinect* cameras per location. Prior to the study, the cameras in each location were registered into the coordinate space of the IR tracking system. At runtime, images were retrieved from the cameras and passed to a reconstruction pipeline. The pipeline was implemented using OpenGL and GLSL, as well as C++ for CPU-GPU transfer, compression, and transmission operations. It begins with a texture processing stage, where the depth maps are downsampled, cleaned and filtered. A downsampling factor of 0.32 was applied for the depth maps, so that avatar frames could be reconstructed and transmitted consistently at 30 Hz. Triangles are generated by joining vertices created by unprojecting neighbouring depth pixels from the processed depth images, provided a maximum depth disparity between neighbouring pixels is not exceeded [68]. For efficient GPU-based outprojection of triangle meshes, we employ compact occupancy tables inspired by those used by the *Marching Cubes* algorithm [41], which specify the triangle configuration to be extracted from a 2×2 depth texel neighbourhood in a branch-free manner.

The *meshoptimizer* library[5] is used to remove redundant vertex data, reduce vertex cache misses [35], and simplify the mesh, resulting in approximately 20k triangles per reconstructed user. Final colour textures were created from the RGB views of each Kinect camera by pre-blending contributions per triangle [37]. The mesh buffers and blended textures are compressed, before being transmitted to the remote participants' Unity applications using the ZeroMQ library[6]. Avatar reconstruction is performed in each location by a PC equipped with an Intel(R) Xeon(R) E5-2687W v4 CPU at 3.00 GHz and an NVIDIA Quadro RTX 6000 GPU.

The shutter glasses that are required for our multi-user projection system are fitted with retroreflective markers to enable pose tracking. Since both the depth and tracking cameras use IR light, the markers cause artefacts in the depth images, which manifest as additional geometry around the head (see Figure 3). To mitigate this effect, a region around the user's tracked head position is defined, where no triangles may be created.

3.5.2 Binaural Audio

Audio was delivered to participants using 3D-printed circumaural open headphones [45] that enabled collocated and remote conversation partners to be heard simultaneously. The headphone volume was adjusted to match the perceived loudness of the remote and collocated users' voices. Voice signals were captured with a Shure WCM16 headworn hypercardioid condenser microphone to minimise crosstalk. Microphone and headphone signals were transmitted via analogue wireless systems (Shure QLXD1/4 and Sennheiser IEM G4) to and from a MOTU M4 audio interface connected to the PC running the Unity application.

Position-dynamic binaural audio was realized with an extended version of the open-source pyBinSim renderer [46], which received remote users' transmitted microphone signals from instances of a Unity Audio Spatializer plugin[7] and local users' tracking data from the ART tracking system, before processing the audio. The extended version of pyBinSim[8] is available online.

The rendering used a set of Head-Related Transfer Functions (HRTFs) and measured room impulse responses of the respective rooms. In this case, the measurements were conducted using a microphone array based on the Spatial Decomposition Method (SDM) introduced by Amengual et al. [18]. The array was positioned at a distance of 2 meters from the centre of the screen. The SDM was further used to create a set of Binaural Room Impulse Responses (BRIRs) in combination with the SADIE II HRTF database (subject D2 - Kemar) [2] with a resolution of 2 degrees.

Direct sound and early reflections up to 64 ms are dynamically updated and scaled relative to the late reverb using the inverse distance law. The late reverb 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 [47]. The DIOT condition was realised in Unity without rendering of distance attenuation or room characteristics. End-to-end audio transmission latency was calculated at 250 ms.

3.5.3 Virtual Environment

The remote participants, visible on the projection screen, were placed in a virtual environment that appeared to be an extension of the viewer's real environment. In particular, the floor of the virtual room matched the real floor; grey in location A (Figure 1d) and wood-coloured in location B (Figure 1c). The width of the virtual room also corresponded to the width of the respective real room.

4 Results

4.1 Data Analysis

This section outlines the analysis approach for each data source.

4.1.1 Questionnaire Analysis

To improve the reliability of the insights obtained from the questionnaire, we analysed the responses in a combinatorial manner. The question groups formed by items 2 to 7 and 8 to 13 are taken

[4] https://www.4players.io/odin/

[5] https://github.com/zeux/meshoptimizer

[6] https://zeromq.org/

[7] https://github.com/vrsys/unity2pybinsim

[8] https://github.com/tuil-emt/PyBinSim_AMR

ID	Question
Overall Experience	
1	How would you rate the overall experience?
Perceived Message Understanding (Collocated)	
2	My thoughts were clear to the CP.
3	The thoughts of the CP were clear to me.
4	It was easy to understand the CP.
5	The CP found it easy to understand me.
6	Understanding the CP was difficult.
7	The CP had difficulty understanding me.
Perceived Message Understanding (Remote)	
8	My thoughts were clear to the RPs.
9	The thoughts of the RPs were clear to me.
10	It was easy to understand the RPs.
11	The RPs found it easy to understand me.
12	Understanding the RPs was difficult.
13	The RPs had difficulty understanding me.
Factor: Group Cohesion	
21	We worked well as a group during the task.
22	The conversation felt natural.
20	I felt like all participants were together as a group.
23	It was easy to know who was speaking.
Factor: Verbal Communication Quality	
17	The RPs sounded realistic.
18	The RPs sounded distracting.
24	We interrupted each other often.
25	It was easier to communicate with the CP than the RPs.
Factor: Visual Quality	
15	The appearance of the RPs was pleasing.
14	It was easy to know where the RPs were looking.
16	The appearance of the RPs was distracting.
Factor: Coherence	
19	The RPs' appearance and voice were of the same quality.
Discomfort	
26	How do you feel compared to when entering the room?

Table 1: Questionnaire items listed in item groups. Collocated Participant (CP) and Remote Participant (RP) are abbreviated in the table, but not in the questionnaire. The ID column reflects the order in which items appeared in the questionnaire.

from the *perceived message understanding* subscale of the validated Networked Minds questionnaire [7], and were combined to calculate perceived message understanding scores relating to collocated and remote participants, respectively. An exploratory factor analysis was conducted on responses to questionnaire items 14 to 25 (items 1 and 26 were excluded). The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis (KMO = 0.83). Horn's parallel analysis [23] indicated that four factors should be used. The four factors accounted for 43.8% of the variance in the dataset. The factor loadings after rotation (using oblique *oblimin* rotation) suggest the following factors, which were tested for reliability using Cronbach's alpha: Group Cohesion ($\alpha = 0.82$), Verbal Communication Quality ($\alpha = 0.62$), Visual Quality ($\alpha = 0.67$), and Coherence (which had only one item and therefore no reliability score). The items that comprise each factor are shown in Table 1. Scores for each item group are calculated by taking the mean of responses to items in the group (after inverting responses to negatively phrased questions). Factor loadings for each item are provided in Appendix A.

In order to perform a factorial ANOVA to determine the effect of the visual and aural spatial fidelity on the clustered questionnaire responses, we apply the aligned rank transform (ART), which allows use of a parametric test with non-normally distributed data [72].

4.1.2 Recorded Speech Analysis

We applied parametric conversation surface structure analysis as described in ITU-T Rec. P.1305 [31], adapted to our 4-party conversation use case. This approach is based on the conversational state model of speech, which represents four-party conversations in 16 states, including *mutual silence*, *single-talk* of any participant and all combinations of *multi-talk*. The state classification is generated from temporal occurrences of speech activity, derived in turn through Voice Activity Detection (VAD). To account for transmission delay, which effectively yields diverging conversational realities for each interlocutor at each site, the state classification approach is applied for each participant's perspective. To facilitate this, we recorded local microphone signals and received signals at each interlocutor's site. After preprocessing of the recordings, including downsampling to 16 kHz and normalisation, we performed VAD with *py-webrtcvad*[9] in frames of 10 ms duration with the aggressiveness mode set to 3. Based on the recommendation by Brady [10], we treat detected talkspurts below 15 ms as silence, and pauses below 200 ms as active speech. After classification into conversation states, we derive the likelihood of each state.

4.1.3 Movement Analysis

Analysis of participants' movements yields objective measures that can reveal changes in communication behaviour under different experimental conditions. In particular, changes in gaze behaviour can indicate the focus of visual attention, and have been linked to conversational engagement [29]. Otsuka [51] investigates gaze in a telepresence scenario, using the proportion of time one participant gazes at another as a proxy for degree of interest and attention, the duration of unbroken gaze as a signifier of depth of focus, and frequency of looks as a proxy for dispersion of attention. We conduct a similar analysis of participants' gaze behaviour by estimating gaze direction from head pose data, and classifying participants' gaze state throughout each trial as *averted* (not looking at another participant), looking at the *collocated*, *remote opposite* or *remote adjacent* participant, or in *transition* between the other states. The proportion of each trial spent in a gaze state is referred to as the *likelihood* of the state. *Glances* are unbroken intervals, during which one of the other participants is the gaze target. We calculate the duration and frequency of glances at each gaze target. *Mutual gaze* likelihood and *mutual glance* metrics are derived by comparing participants' gaze states. The gaze metrics are calculated on a per-trial basis for each user. Per-user results are aggregated to obtain a single value per trial for each group, either by averaging (for likelihood and glance duration) or adding (glance frequencies).

Gaze states are derived from head transform data at a fixed sample rate of 10 Hz, using a simple geometric model. A transition is marked if the moving average of angular rotation speed exceeds a threshold (30 deg/s). If no transition is detected, the angle θ between the head's forward vector $\mathbf{v_e}$ (i.e. the viewing direction) and the vector from the head to the head of each other participant $\mathbf{v_p}$ is calculated (see Figure 2). If $\theta < 25$ deg, gaze state is set accordingly with that participant as the target; if not, the gaze state is set as *averted*. Comparisons of gaze-metric means were performed with paired samples t-tests, after confirming normality with a Shapiro-Wilk test, unless specified.

4.2 Effect of Visual and Aural Realism

4.2.1 Questionnaire Responses

Analysis of the questionnaire results (shown in Table 2) detected no significant main effect of aural spatial realism on the mean score of any item group. A positive effect of visual realism was observed on

[9] https://github.com/wiseman/py-webrtcvad

		Visual Realism					Aural Realism		
Questionnaire Item Group	Mean	VOLU	ABST	p	sig.	$F_{1,189}$	SPAT	DIOT	p
Overall Experience	5.89	**5.98**	5.80	0.107		2.6	**5.92**	5.86	0.647
Perceived Message Understanding for CP	6.20	**6.20**	6.19	0.827		0.0	**6.21**	6.19	0.393
Perceived Message Understanding for RP	5.99	**6.09**	5.89	0.005	**	8.1	5.97	**6.00**	0.906
Group Cohesion	6.17	**6.37**	5.98	<.001	***	26.8	6.17	**6.18**	0.541
Verbal Communication Quality	5.04	**5.13**	4.95	0.052		3.8	5.03	**5.04**	0.697
Visual Fidelity	4.68	**4.90**	4.46	<.001	***	13.4	**4.74**	4.63	0.422
Coherence	4.23	4.05	**4.41**	0.024	*	5.2	4.23	4.23	0.973
Simulator Sickness	6.45	**6.52**	6.38	0.121		2.4	6.44	**6.46**	0.765

Table 2: Mean scores and significant effects for each item group. The F value is omitted for main effect of aural realism as no significant effects were observed. Statistical significance (sig.) marked as: * <= .05; ** <= .01; *** <= .001.

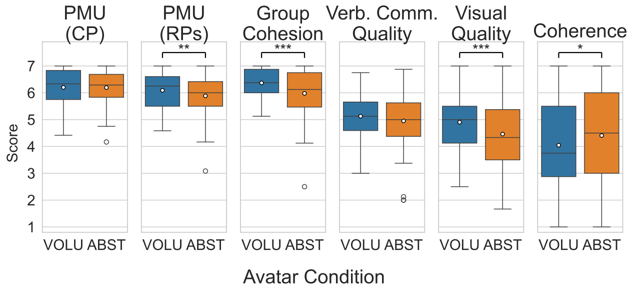


Figure 4: Effect of visual realism on item group response distributions. PMU = Perceived Message Understanding.

perceived message understanding with remote participants (RPs), group cohesion, and visual quality scores (Figure 4). Analysis of the coherence score indicates that participants rated quality of RPs' appearance and voice as more similar under the ABST condition.

4.2.2 Gaze Analysis

Analysis of gaze metrics (Table 3) show that under the VOLU condition, more time was spent looking at and sharing gaze with RPs, while glances at the remote adjacent user were significantly longer. Higher glance frequency at (and mutual glances with) the remote adjacent user also occurred under the VOLU condition. Gaze metrics are displayed for each gaze target in Figures 5, 6, and 7.

Participants spent more time looking at their collocated teammate under the DIOT audio condition than the SPAT condition ($t(15) = 2.2$, $p = .045$). In the DIOT condition, more mutual glances with the remote opposite participants were observed (as shown by a Wilcoxon Signed-Rank test, $Z = 26.0$, $p = .029$).

4.2.3 Conversation Analysis

The likelihood of mutual silence, single talk, and multi-talk states for each experimental condition is shown in Figure 8. Silence and single talk are the most common states, even when compared to all multi-talk states combined. No significant differences between the likelihood of the states under different conditions were observed.

4.3 Comparing Collocated and Remote Communication

4.3.1 Questionnaire Responses

The two questionnaire item groups relating to perceived message understanding ask participants to rate the same characteristics of the communication for collocated and remote participants respectively (see Table 1). The comparison of understanding scores of collocated communication (mean = 6.20) and remote communication (mean = 5.99) shows a significant positive impact of collocation, indicated by the Wilcoxon Signed-Rank test ($Z = 582$, $p = .022$).

Metric	Mean	VOLU	ABST	p	sig.	$t(15)$
Gaze Likelihood (%)						
Collocated	9.70	9.35	**10.05**	0.165		-1.5
Rem. Adj.	12.04	**14.29**	9.80	<.001	***	6.0
Rem. Opp.	37.28	**37.88**	36.68	0.043	*	2.2
Collocated (Mut.)	1.66	1.66	1.66	0.973		-0.0
Rem. Adj. (Mut.)	1.68	**2.23**	1.14	<.001	***	4.3
Rem. Opp. (Mut.)	13.50	**14.21**	12.79	0.023	*	2.5
Glance Duration (s)						
Collocated	1.63	1.54	**1.72**	0.059		-2.0
Rem. Adj.	1.47	**1.60**	1.35	<.001	***	4.1
Rem. Opp.	1.78	1.76	**1.81**	0.336		-1.0
Glance Frequency (glances per min)						
Collocated	14.08	**14.30**	13.85	0.527		0.6
Rem. Adj.	19.49	**21.44**	17.54	0.001	***	3.9
Rem. Opp.	52.54	**53.45**	51.64	0.205		1.3
Collocated (Mut.)	2.34	**2.48**	2.20	0.228		1.3
Rem. Adj. (Mut.)	2.89	**3.56**	2.22	0.002	**	3.6
Rem. Opp. (Mut.)	19.48	**20.05**	18.90	0.168		1.4

Table 3: Effect of avatar realism conditions on gaze metrics.

4.3.2 Conversation Analysis

No significant differences were found between the likelihood of double talk of collocated pairs (mean = 3.39%), remote adjacent pairs (mean = 3.27%), and remote opposite pairs (mean = 3.24%).

4.3.3 Gaze Analysis

Comparison of overall gaze likelihoods show that the most common gaze target is the remote opposite participant (see Figure 9). Slightly more time is spent looking at the remote adjacent user than the collocated user, but the difference is not statistically significant. Mutual gaze likelihood also shows that more mutual gaze occurred with the remote opposite participant, while similar amounts of mutual gaze occurred with the collocated and remote adjacent participants. Similarly, the target with the highest frequency of glances is the remote opposite participant, while the frequency of glances at the remote adjacent participant was lower, but still higher than the frequency of glances at the collocated participant. The difference in glance frequency between remote adjacent and collocated is significant ($t(15) = 3.03$, $p = .008$) The mean duration of glances at (and mutual glances with) the remote adjacent participant is lower than at the other participants, but not significantly.

5 Discussion

5.1 RQ1: Effect of Visual Realism

We observed various positive effects of visual realism on communication in group-to-group telepresence. The positive impact on ratings of perceived message understanding, group cohesion, and

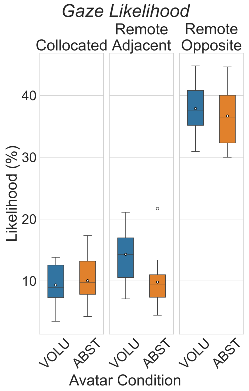


Figure 5: Gaze likelihood by avatar condition.

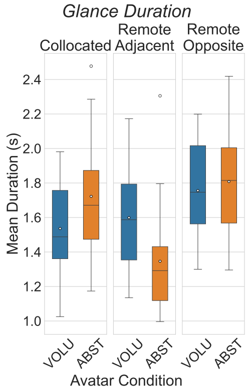


Figure 6: Glance duration by avatar condition.

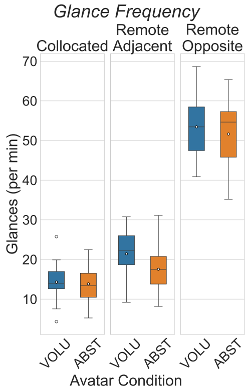


Figure 7: Glance frequency by avatar condition.

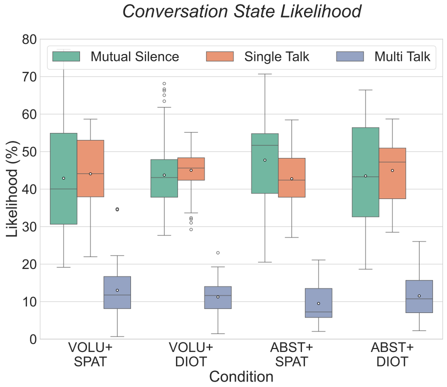


Figure 8: The likelihood of conversational states under each experimental condition.

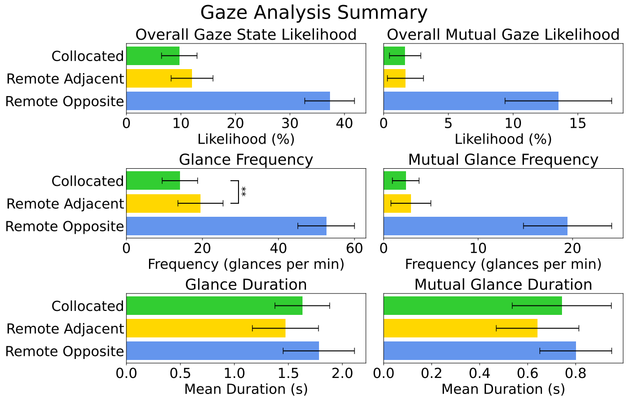


Figure 9: Overall gaze analysis metrics. Significance only shown for collocated/remote adjacent comparisons.

visual quality is likely to be explained by the more expressive nature of the volumetric avatar type. In particular, the visibility of facial expressions is likely to be an important factor, as they are not represented at all in the abstract avatar case. Facial expressions can also explain users' reported increase in the ability to tell who was talking in the volumetric case (part of the visual quality score). These results can be explained in the context of Media Richness Theory [14], which states that rich media (e.g. media that utilise many cues and channels) can reduce ambiguity and are therefore suited to solving complex tasks. In our study, the cognitive load of converging on an agreed solution [36] may be compounded by confusion about who is speaking or being addressed. Therefore, avatars that convey richer cues can reduce cognitive load, influencing participants' ratings of mutual understanding and group cohesion.

The higher scores for measures related to communication quality for volumetric avatars over abstract avatars in a group-to-group setting aligns with results from previous studies of one-to-one telepresence scenarios [74, 13, 16], despite the ability to compare virtual stimuli directly with real stimuli.

The more realistic avatar condition also had an impact on several gaze metrics. More time was spent looking at and sharing mutual gaze with RPs, and the frequency of glances at RPs increased. These changes, especially when considered in the context of the questionnaire responses, suggest that participants paid more attention to RPs and were more engaged in the conversation under the VOLU condition. The expressiveness of the realistic avatar can also explain these results; when more visual cues are conveyed, one is more likely to observe the avatar to interpret those cues.

No effect of visual realism could be observed from a comparison of conversational state likelihoods. Given the positive subjective ratings of the volumetric avatar case, one might expect conversation states to reflect higher interactivity. The low proportion of the conversation during which missing visual cues might affect the conversation structure could mask any impact on state likelihood metrics. Accordingly, an in-depth analysis of speaker transitions characterised by walks of states could provide more information.

We can conclude that, even when users can compare virtual, visual stimuli against real stimuli, it is still beneficial to represent remote users with a realistic avatar that can convey more non-verbal cues. Additionally, we note that a significant positive effect of avatar realism on communication is possible even when the gaze information is not conveyed.

5.2 RQ2: Effect of Aural Realism

We identified very few significant effects of aural realism. This is somewhat surprising, as we expected the complex group-to-group communication scenario with multiple audio sources to increase the utility of spatial audio. For example, when multiple people are speaking in the ABST avatar conditions, in which lack of facial expressions means that the avatars do not provide a clear visual indication of who is speaking, one would expect spatial audio to help participants to identify the speaker (leading to effects on group cohesion scores), or lead to a reduction in mean glance duration and increased glance frequency as participants scanned to identify the speaker. The increase in time spent looking at the collocated user under the DIOT condition (see Section 4.2.2) could reflect an orienting reflex where participants automatically look in the direction of the sound source [52], since the collocated user's voice was the only localised sound source under that condition.

The reason for the absence of significant differences between the spatial and diotic conditions may be hinted at by the scores obtained for verbal communication quality for each audio condition. The relevant item group includes items that elicit direct ratings of characteristics of RPs' voice signals (Q17 and Q18). The lack of a significant effect of the audio condition on the verbal communication quality suggests that the difference in realism between the conditions was not sufficient for participants to notice the change; while the difference in visual realism conditions was much clearer.

It is also possible that spatial cues provided by audio stimuli were redundant. Listeners may have associated voices with avatars, making spatial cues redundant and therefore inhibiting their processing [9]. Participants could not move freely because of the fixed display, meaning that the need for sound source localisation was limited. In more acoustically complex scenes with free movement (e.g. if using HMDs) as well as more coexisting sound sources, spatial audio could be more important in locating and understanding interlocutors, for example, during dynamic group formation and collaboration in crowded VEs. A recent study [24] found that in a triadic communication scenario with free movement, spatial audio had no impact on holistic subjective measures of communication, hinting that more than two audio sources in acoustically complex scenes are required to produce measurable behavioural changes.

5.3 RQ3: Collocated and Remote Communication

Participants reported higher perceived message understanding when communicating with the collocated participant, compared to the remote participants. This is unsurprising, given that the collocated participants were communicating in an unmediated manner, while communication with the RPs is subject to the limitations of the medium (the telepresence system).

Analysis of the gaze behaviour does not unambiguously reflect the subjective differences. Firstly, all gaze metrics show a bias towards the remote opposite user, who receives more visual attention and more frequent glances. This can be explained by the positioning of the participants on the corners of a square, facing towards the centre; therefore facing towards the remote opposite participant, and meaning that more gaze interactions with that participant were inevitable. A comparable effect was also observed in another telepresence study with a similar participant layout [51]. It makes more sense, then, to compare gaze interactions with the collocated and remote adjacent participants, which are less likely to be affected by this positional bias. Participants made significantly more glances at the remote adjacent participant than the collocated participant, while some other metrics (e.g. overall gaze likelihood) showed a non-significant tendency in favour of the remote participant. If we take glance frequency as an indicator of focus, then it seems participants were more focused on the remote participant; perhaps as a result of increased difficulty in interpreting non-verbal communication cues when compared to the collocated participant. The visual characteristics of the avatars are certainly not at the level of cues received from the collocated person. Furthermore, the visual acuity achieved by the 4k projection system – interacting also with the camera quality in the VOLU condition – is certainly much lower than what the human eye can resolve.

5.4 Limitations

Our study relied on two large multi-user projection screen systems that are installed in different physical locations. This made some desirable qualities of a mediated communication study, including mixing of collocated pairs, and the presence of a face-to-face reference, infeasible. The number of participants also meant that manual coding of recorded trials, which would have produced more accurate gaze state classification and richer gesture data, has not yet been undertaken due to time constraints. We also note that the test environments were not acoustically treated and were subject to the noise of the equipment, which may have affected the hearing conditions via the open headphones.

The multi-user projection screens have an effective pixel density of 16 to 19 PPD (based on the user positions shown in Figure 2), which is lower than that of modern HMDs (e.g. the Meta Quest 3 has 25 PPD). The projection systems' motion-to-photon latency was measured at 140 ms, higher than would be expected from HMDs, which use warping techniques to achieve very low motion-to-photon latency. However, we do not believe that the findings of this study were significantly affected by these factors. Instead, we posit that higher visual fidelity would be likely to emphasise the positive effects of visual realism by more clearly conveying the non-verbal cues in the volumetric avatar condition. The high motion-to-photon latency may have degraded the users' overall experience, but is unlikely to disproportionally affect any of the study conditions. Regarding spatial audio, we note that since tracking data was sent directly from the tracking system to the binaural rendering server, the spatial audio latency was lower than the motion-to-photon latency, conforming to suggested latency thresholds [58].

Volumetric avatar data was transmitted over the national high-speed research network (DFN). Available network bandwidth is likely to be lower in practical applications of volumetric avatars for communication (e.g. when data is sent wirelessly to mobile clients). However, new compression techniques and standards [69], along with the faster mobile data networks, will enable streaming of high-quality volumetric avatars for telepresence applications.

6 Conclusion and Future Work

In this work, we have investigated how communication is affected by the combination of real and virtual user representations that occurs in group-to-group telepresence. In our study, two groups of two collocated participants communicated between two remote locations using a telepresence system. We have found that higher visual realism of remote users' avatars increases perceived message understanding and group cohesion, while behavioural analysis indicates that more visual attention is paid to remote interlocutors when they are represented by realistic avatars. Despite the presence of multiple audio sources, we did not find that applying spatial processing to users' voices had a noticeable effect on communication.

When comparing communication with collocated users against communication with remote users, we found that unmediated communication with collocated users was rated as more effective. Simultaneously, behavioural analysis showed that more visual attention was paid to remote users, potentially because more effort was needed to decode the non-verbal cues conveyed by the avatars. This suggests scope for improvement of the virtual user representations to match the fidelity of real users. This is likely to be dependent on the ability of the avatars to convey gaze signals, the resolution of volumetric avatars, and the fidelity of the display system.

In future work, a similar comparison of collocated and remote communication in a group-to-group telepresence system based on HMDs would be relevant, given the increasing prevalence of such devices in MR systems. Fully immersive HMD-based telepresence systems with multiple users may elicit measurable effects of spatial audio, since virtual audio sources can be behind the user, increasing the utility of spatialisation in identifying the speaker. Modern HMDs support capture of facial expressions and gaze tracking; a comparison of collocated and remote communication with avatars animated using such information would indicate whether animated avatars would be preferred over volumetric avatars that convey the real appearance of collaboration partners, including their HMDs, in a group-to-group telepresence setting. Moreover, a comprehensive examination of different levels and dimensions (e.g. avatar reconstruction rate and level of detail) and their effect on communication would provide valuable guidance to designers of MR systems.

Our study provides a reference for future behavioural analyses of multi-party communication between remote groups. Nevertheless, extensive research is still needed to optimally integrate remote users with collocated groups as equal communication partners.

Acknowledgments

This work is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under the project ID 444532506, SPP2236 - AUDICTIVE - Auditory Cognition in Interactive Virtual Environments. We thank our respective research groups at Bauhaus-Universität Weimar and TU Ilmenau for their support.

References

[1] Advanced Realtime Tracking GmbH & Co. KG. ART. Accessed 09.17.2024. 5

[2] 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 5

[3] S. Aseeri and V. Interrante. The Influence of Avatar Representation on Interpersonal Communication in Virtual Social Environments. *IEEE TVCG*, 27(5):2608–2617, 2021. doi: 10.1109/TVCG.2021.3067783 2

[4] J. J. Baldis. Effects of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences. In *CHI '01*, pp. 166–173. ACM, NY, USA, 2001. doi: 10.1145/365024.365092 2, 3

[5] S. Beck, A. Kunert, A. Kulik, and B. Froehlich. Immersive group-to-group telepresence. *IEEE TVCG*, 19(4):616–625, Apr. 2013. doi: 10.1109/TVCG.2013.33 1, 2

[6] R. Bednarik, S. Eivazi, and M. Hradis. Gaze and conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement. In *Proc. of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, Gaze-In '12. ACM, New York, NY, USA, 2012. doi: 10.1145/2401836.2401846 3

[7] F. Biocca, C. Harms, and J. Gregg. The Networked Minds Measure of Social Presence: Pilot Test of the Factor Structure and Concurrent Validity. *4th Int. Workshop on Presence, Philadelphia*, 01 2001. 4, 6

[8] J. Blascovich, J. Loomis, A. C. Beall, K. R. Swinth, C. L. Hoyt, and J. N. Bailenson. Immersive Virtual Environment Technology as a Methodological Tool for Social Psychology. *Psychological Inquiry*, 13(2):103–124, 2002. 2

[9] 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 8

[10] P. T. Brady. A Statistical Analysis of On-Off Patterns in 16 Conversations. *Bell System Technical Journal*, 47(1):73–91, 1968. doi: 10.1002/j.1538-7305.1968.tb00031.x 3, 6

[11] W. A. S. Buxton. Telepresence: integrating shared task and person spaces. In *Proc. of the Conf. on Graphics Interface '92*, pp. 123–129. Morgan Kaufmann Publishers Inc, San Francisco, USA, 1992. 1, 2

[12] C. Cao, T. Simon, J. K. Kim, G. Schwartz, M. Zollhoefer, S.-S. Saito, S. Lombardi, S.-E. Wei, D. Belko, S.-I. Yu, Y. Sheikh, and J. Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 41(4), jul 2022. doi: 10.1145/3528223.3530143 2

[13] S. Cho, S.-w. Kim, J. Lee, J. Ahn, and J. Han. Effects of volumetric capture avatars on social presence in immersive virtual environments. In *2020 IEEE Conf. on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2020. doi: 10.1109/vr46266.2020.00020 2, 8

[14] R. L. Daft and R. H. Lengel. Organizational Information Requirements, Media Richness and Structural Design. *Management Science*, 32(5):554–571, 1986. doi: 10.1287/mnsc.32.5.554 8

[15] B. Ens, J. Lanir, A. Tang, S. Bateman, G. Lee, T. Piumsomboon, and M. Billinghurst. Revisiting collaboration through mixed reality: The evolution of groupware. *Int. Journal of Human-Computer Studies*, 131:81–98, 2019. doi: 10.1016/j.ijhcs.2019.05.011 1

[16] G. Gamelin, A. Chellali, S. Cheikh, A. Ricca, C. Dumas, and S. Otmane. Point-cloud avatars to improve spatial communication in immersive collaborative virtual environments. *Personal Ubiquitous Comp.*, 25(3):467–484, 2020. doi: 10.1007/s00779-020-01431-1 2, 8

[17] M. Garcia and A. Raake. Impairment-factor-based audio-visual quality model for IPTV. In *2009 Int. Workshop on Quality of Multimedia Experience*, pp. 1–6, 2009. doi: 10.1109/QOMEX.2009.5246985 3

[18] S. V. A. Gari, J. Arend, P. Calamia, and P. Robinson. Optimizations of the Spatial Decomposition Method for Binaural Reproduction. *Journal of the Audio Engineering Society*, 68(12):959 – 976, 12 2020. doi: 10.17743/jaes.2020.0063 5

[19] M. Gross, S. Würmlin, M. Naef, E. Lamboray, C. Spagno, A. Kunz, E. Koller-Meier, T. Svoboda, L. Van Gool, S. Lang, K. Strehlke, A. Vande Moere, and O. Staadt. Blue-c: A Spatially Immersive Display and 3D Video Portal for Telepresence. *ACM Trans. Graph.*, 22(3):819–827, jul 2003. doi: 10.1145/882262.882350 2

[20] E. T. Hall. The hidden dimension: an anthropologist examines man's use of space in public and private. 1969. 3

[21] A. T. Hayes, C. E. Hughes, and J. Bailenson. Identifying and Coding Behavioral Indicators of Social Presence With a Social Presence Behavioral Coding System. *Frontiers in Virtual Reality*, 3, 2022. doi: 10.3389/frvir.2022.773448 3

[22] 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 3

[23] J. L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965. doi: 10.1007/BF02289447 6

[24] F. Immohr, G. Rendle, C. Kehling, A. Lammert, S. Göring, B. Froehlich, and A. Raake. Subjective Evaluation of the Impact of Spatial Audio on Triadic Communication in Virtual Reality. In *QoMEX '24*, pp. 262–265, 2024. doi: 10.1109/QoMEX61742.2024.10598292 2, 3, 9

[25] F. Immohr, G. Rendle, A. Lammert, A. Neidhardt, V. M. zur Heyde, B. Froehlich, and A. Raake. Evaluating the Effect of Binaural Auralization on Audiovisual Plausibility and Communication Behavior in Virtual Reality. In *IEEE VR '24*, pp. 849–858, 2024. doi: 10.1109/VR58804.2024.00104 2, 3

[26] F. Immohr, G. Rendle, A. Neidhardt, S. Göring, R. R. Ramachandra Rao, S. Arevalo 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 2023 ACM Int. Conf. on Interactive Media Experiences*, IMX '23, pp. 209–215. ACM, New York, NY, USA, 2023. doi: 10.1145/3573381.3596458 2, 3

[27] A. Irlitti, M. Latifoglu, T. Hoang, B. V. Syiem, and F. Vetere. Volumetric Hybrid Workspaces: Interactions with Objects in Remote and Co-located Telepresence. In *Proc. of the CHI Conf. on Human Factors in Computing Systems*, pp. 1–16. ACM, New York, NY, USA, 2024. doi: 10.1145/3613904.3642814 1, 2

[28] R. Ishii and Y. I. Nakano. An empirical study of eye-gaze behaviors. In *Proc. of the 2010 workshop on Eye gaze in intelligent human machine interaction*, pp. 33–40. ACM, New York, NY, USA, 2010. doi: 10.1145/2002333.2002339 3

[29] R. Ishii, Y. I. Nakano, and T. Nishida. Gaze awareness in conversational agents. *ACM Trans. on Interactive Intelligent Systems*, 3(2):1–25, 2013. doi: 10.1145/2499474.2499480 6

[30] ITU-T Rec. P.1301. Subjective quality evaluation of audio and audiovisual multiparty telemeetings, 2017. 3, 4

[31] ITU-T Rec. P.1305. Effect of delays on telemeeting quality, 2016. 6

[32] 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 2, 3

[33] S. Kang, G. Kim, K.-T. Lee, and S. Kim. How Do Background and Remote User Representations Affect Social Telepresence in Remote Collaboration?: A Study with Portal Display, a Head Pose-Responsive Video Teleconferencing System. *Electronics*, 12(20), 2023. doi: 10.3390/electronics12204339 2

[34] P. Kauff and O. Schreer. An Immersive 3D Video-Conferencing System Using Shared Virtual Team User Environments. In *CVE '02*, pp. 105–112. ACM, New York, NY, USA, 2002. doi: 10.1145/571878.571895 1, 2

[35] B. Kerbl, M. Kenzel, E. Ivanchenko, D. Schmalstieg, and M. Steinberger. Revisiting The Vertex Cache: Understanding and Optimizing Vertex Processing on the modern GPU. *Proc. ACM Comput. Graph. Interact. Tech.*, 1(2), aug 2018. doi: 10.1145/3233302 5

[36] G. L. Kolfschoten and F. Brazier. Cognitive Load in Collaboration–Convergence. In *2012 45th Hawaii Int. Conf. on System Sciences*, pp. 129–138, 2012. doi: 10.1109/HICSS.2012.156 8

[37] A. Kreskowski, S. Beck, and B. Froehlich. Output-Sensitive Avatar Representations for Immersive Telepresence. *IEEE TVCG*, 28(7):2697–2709, 2022. 5

[38] A. Kulik, A. Kunert, S. Beck, R. Reichel, R. Blach, A. Zink, and B. Froehlich. C1x6: a stereoscopic six-user display for co-located collaboration in shared virtual environments. *ACM Trans. Graph.*, 30(6):1–12, 2011. doi: 10.1145/2070781.2024222 2

[39] A. Lammert, G. Rendle, F. Immohr, A. Neidhardt, K. Brandenburg,

A. Raake, and B. Froehlich. Immersive Study Analyzer: Collaborative Immersive Analysis of Recorded Social VR Studies. *IEEE TVCG*, 30(11):7214–7224, 2024. doi: 10.1109/TVCG.2024.3456146 5

[40] J. Lawrence, D. Goldman, S. Achar, G. M. Blascovich, J. G. Desloge, T. Fortes, E. M. Gomez, S. Häberling, H. Hoppe, A. Huibers, C. Knaus, B. Kuschak, R. Martin-Brualla, H. Nover, A. I. Russell, S. M. Seitz, and K. Tong. Project starline: a high-fidelity telepresence system. *Proc. SIGGRAPH Asia*, 40(6), 2021. doi: 10.1145/3478513.3480490 2

[41] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. *SIGGRAPH Comput. Graph.*, 21(4):163–169, aug 1987. doi: 10.1145/37402.37422 5

[42] A. Maimone and H. Fuchs. Real-time volumetric 3D capture of room-sized scenes for telepresence. In *2012 3DTV-Conf.: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, pp. 1–4, 2012. doi: 10.1109/3DTV.2012.6365430 2

[43] M. R. Miller, N. Sonalkar, A. Mabogunje, L. Leifer, and J. Bailenson. Synchrony within Triads using Virtual Reality. *Proc. of the ACM on Human-Computer Interaction*, 5, 2021. doi: 10.1145/3479544 3

[44] J. Mulligan, X. Zabulis, N. Kelshikar, and K. Daniilidis. Stereo-based environment scanning for immersive telepresence. *IEEE Trans. on Circuits and Systems for Video Technology*, 14(3):304–320, 2004. doi: 10.1109/TCSVT.2004.823390 2

[45] A. Mülleder and F. Zotter. Ultralight circumaural open headphones. *journal of the audio engineering society*, (102), may 2023. 5

[46] A. Neidhardt, F. Klein, N. Knoop, and T. Köllmer. Flexible python tool for dynamic binaural synthesis applications. In *Audio Engineering Society Conv. 142*. Audio Engineering Society, 2017. 5

[47] 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 Conv., Milan, Italy*, 2018. 5

[48] 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 *Proc. of the 2nd Annual Meeting of the Symp. on Human-Computer Interaction for Work*, pp. 1–10. ACM, New York, NY, USA, 2023. doi: 10.1145/3596671.3598578 2, 3

[49] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, V. Tankovich, C. Loop, Q. Cai, P. A. Chou, S. Mennicken, J. Valentin, V. Pradeep, S. Wang, S. B. Kang, P. Kohli, Y. Lutchyn, C. Keskin, and S. Izadi. Holoportation. In *UIST '16*, pp. 741–754. ACM, New York, NY, USA, 2016. doi: 10.1145/2984511.2984517 2

[50] K. Otsuka. MMSpace: Kinetically-augmented telepresence for small group-to-group conversations. In *2016 IEEE Virtual Reality (VR)*, pp. 19–28, 2016. doi: 10.1109/VR.2016.7504684 1

[51] K. Otsuka. Behavioral Analysis of Kinetic Telepresence for Small Symmetric Group-to-Group Meetings. *IEEE Trans. on Multimedia*, 20(6):1432–1447, 2018. doi: 10.1109/TMM.2017.2771396 1, 3, 6, 9

[52] I. P. Pavlov. Conditioned Reflexes. 1927. 8

[53] T. Pejsa, J. Kantor, H. Benko, E. Ofek, and A. Wilson. Room2Room. In *CSCW '16*, pp. 1716–1725. ACM, New York, NY, USA, 2016. doi: 10.1145/2818048.2819965 2

[54] T. Potter, Z. Cvetkovic, and E. de SENA. On the Relative Importance of Visual and Spatial Audio Rendering on VR Immersion. *Frontiers in Signal Processing*, 2, 2022. doi: 10.3389/frsip.2022.904866 3

[55] A. Raake, J. Ahrens, M. Geier, and C. Schlegel. Listening and Conversational Quality of Spatial Audio Conferencing. In *40th International AES Conference*, Oct 2010. 2, 3

[56] G. Rendle, A. Kreskowski, and B. Froehlich. Volumetric Avatar Reconstruction with Spatio-Temporally Offset RGBD Cameras. In *2023 IEEE Conf. Virtual Reality and 3D User Interfaces (VR)*, pp. 72–82, 2023. doi: 10.1109/VR55154.2023.00023 2

[57] 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. 3

[58] N. Sankaran, J. Hillis, M. Zannoli, and R. Mehra. Perceptual thresholds of spatial audio update latency in virtual auditory and audiovisual environments. *The Journal of the Acoustical Society of America*, 140(4_Supplement):3008, 2016. doi: 10.1121/1.4969329 9

[59] S. Scherer, M. Glodek, G. Layher, M. Schels, M. Schmidt, T. Brosch, S. Tschechne, F. Schwenker, H. Neumann, and G. Palm. A generic framework for the inference of user states in human computer interaction: How patterns of low level behavioral cues support complex user states in HCI. *Journal on Multimodal User Interfaces*, 6:117–141, 2012. doi: 10.1007/s12193-012-0093-9 3

[60] 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 3

[61] 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 3

[62] J. Skowronek. *Quality of experience of multiparty conferencing and telemeeting systems*. PhD thesis, 2017. doi: 10.14279/DEPOSITONCE-5811 3

[63] 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 2

[64] W. Standaert, S. Muylle, and A. Basu. An empirical study of the effectiveness of telepresence as a business meeting mode. *Information Technology and Management*, 17(4):323–339, 2016. doi: 10.1007/s10799-015-0221-9 1

[65] Steed, Anthony and Schroeder, Ralph. Collaboration in Immersive and Non-immersive Virtual Environments. In *Immersed in Media: Telepresence Theory, Measurement & Technology*, pp. 263–282. Springer International Publishing, Cham, 2015. doi: 10.1007/978-3-319-10190-3 2

[66] Tivian XI GmbH. Unipark. Accessed 08.04.2024. 4

[67] H. Tu, R. Shao, X. Dong, S. Zheng, H. Zhang, L. Chen, M. Wang, W. Li, S. Ma, S. Zhang, B. Zhou, and Y. Liu. Tele-Aloha: A Telepresence System with Low-budget and High-authenticity Using Sparse RGB Cameras. SIGGRAPH '24. ACM, New York, NY, USA, 2024. doi: 10.1145/3641519.3657491 2

[68] G. Turk and M. Levoy. Zippered polygon meshes from range images. In *Proc. of the 21st Annual Conf. on Computer Graphics and Interactive Techniques*, SIGGRAPH '94, p. 311–318. ACM, New York, NY, USA, 1994. doi: 10.1145/192161.192241 5

[69] V. K. M. Vadakital, A. Dziembowski, G. Lafruit, F. Thudor, G. Lee, and P. R. Alface. The MPEG Immersive Video Standard—Current Status and Future Outlook. *IEEE MultiMedia*, 29(3):101–111, 2022. doi: 10.1109/MMUL.2022.3175654 9

[70] Willem Standaert, Steve Muylle, and Amit Basu. How shall we meet? Understanding the importance of meeting mode capabilities for different meeting objectives. *Information & Management*, 58(1):103393, 2021. doi: 10.1016/j.im.2020.103393 1

[71] J. R. Williamson, J. O'Hagan, J. A. Guerra-Gomez, J. H. Williamson, P. Cesar, and D. A. Shamma. Digital Proxemics: Designing Social and Collaborative Interaction in Virtual Environments. In *CHI '22*. ACM, New York, NY, USA, 2022. doi: 10.1145/3491102.3517594 3

[72] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *CHI '11*, p. 143–146. ACM, New York, NY, USA, 2011. doi: 10.1145/1978942.1978963 6

[73] J. W. Woodworth, N. G. Lipari, and C. W. Borst. Evaluating Teacher Avatar Appearances in Educational VR. In *2019 IEEE Conf. on Virtual Reality and 3D User Interfaces (VR)*, pp. 1235–1236, 2019. doi: 10.1109/VR.2019.8798318 2

[74] K. Yu, G. Gorbachev, U. Eck, F. Pankratz, N. Navab, and D. Roth. Avatars for Teleconsultation: Effects of Avatar Embodiment Techniques on User Perception in 3D Asymmetric Telepresence. *IEEE TVCG*, 27(11):4129–4139, 2021. doi: 10.1109/TVCG.2021.3106480 2, 8

[75] C. Zhang, Q. Cai, P. Chou, Z. Zhang, and R. Martin-Brualla. Viewport: A Distributed, Immersive Teleconferencing System with Infrared Dot Pattern. *IEEE MultiMedia*, 20(1):17–27, 2013. doi: 10.1109/MMUL.2013.12 1