Journal on Multimodal User Interfaces
https://doi.org/10.1007/s12193-020-00331-1

ORIGINAL PAPER

The effects of spatial auditory and visual cues on mixed reality remote collaboration

Jing Yang[1] · Prasanth Sasikumar[2] · Huidong Bai[2] · Amit Barde[2] · Gábor Sörös[3] · Mark Billinghurst[2]

Received: 2 February 2020 / Accepted: 6 June 2020

Abstract
Collaborative Mixed Reality (MR) technologies enable remote people to work together by sharing communication cues intrinsic to face-to-face conversations, such as eye gaze and hand gestures. While the role of visual cues has been investigated in many collaborative MR systems, the use of spatial auditory cues remains underexplored. In this paper, we present an MR remote collaboration system that shares both spatial auditory and visual cues between collaborators to help them complete a search task. Through two user studies in a large office, we found that compared to non-spatialized audio, the spatialized remote expert's voice and auditory beacons enabled local workers to find small occluded objects with significantly stronger spatial perception. We also found that while the spatial auditory cues could indicate the spatial layout and a general direction to search for the target object, visual head frustum and hand gestures intuitively demonstrated the remote expert's movements and the position of the target. Integrating visual cues (especially the head frustum) with the spatial auditory cues significantly improved the local worker's task performance, social presence, and spatial perception of the environment.

Keywords Mixed reality · Augmented reality · Virtual reality · Remote collaboration · Spatial audio · Hand gesture

1 Introduction

This paper explores the impact of using spatial auditory and visual cues in a Mixed Reality (MR) interface for remote collaboration. Remote collaboration enables spatially distant people to work together, which can increase collaborators' productivity and is a cost-effective alternative to business travel and expert visits. However, most people still prefer direct face-to-face communication over current audio-video conferencing solutions, partly because the latter usually fail to convey the implicit non-verbal cues that play an important role in face-to-face collaborations.

The limitations of video conferencing can be addressed by using MR technologies, which seamlessly combine virtual contents with the real environment. MR remote collaboration systems with Head-Mounted Displays (HMDs) can transfer various spatial cues used in face-to-face communications. For example, besides talking like in a phone call [12], users can also share their eye gaze and hand gestures with each other in the form of visual augmentation on the display [22,40]. The local worker's environment can be live captured and streamed to the remote side in 3D [41]. By sharing such communication cues, a worker in a local space can better execute tasks with help from an expert in a remote location.

So far, most MR remote collaboration systems mainly explore the use of visual channel to deliver cues and text information [13,45]. Despite the fact that some non-spatial audio (e.g. verbal communication) is also used, the spatial auditory cues generally remain underexplored in MR remote collaboration, although they have promising effects

✉ Jing Yang
jing.yang@inf.ethz.ch

Prasanth Sasikumar
psas598@aucklanduni.ac.nz

Huidong Bai
huidong.bai@auckland.ac.nz

Amit Barde
amit.barde@auckland.ac.nz

Gábor Sörös
gabor.soros@nokia-bell-labs.com

Mark Billinghurst
mark.billinghurst@auckland.ac.nz

1 Department of Computer Science, ETH Zurich, Zurich, Switzerland

2 Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand

3 Nokia Bell Labs, Budapest, Hungary

Published online: 28 June 2020

and potential benefits. Indeed, some researchers have proposed to convey information using spatial auditory cues in Augmented Reality (AR) or Human-Computer Interaction (HCI) projects [9,20,35,43,49]. In our daily life, we also experience our surroundings via natural auditory perception, with which we can gauge the positions of objects even if they are outside our field of view (FoV). Auditory cues can also be especially useful if the visual sense is occupied or unsafe to use (e.g. while driving) [46], or if showing a large amount of visual information causes a heavy cognitive load [46,48]. Technically, it is already possible to integrate the illusion of natural auditory perception into commercially available AR and Virtual Reality (VR) devices.

In this paper, we present a MR remote collaboration system that features both spatial auditory and visual cues, and we conducted two user studies. Study 1 focuses on the exploration of various auditory cues. Study 2 intends to extend our understanding of hybrid models that include both spatial auditory and visual cues. For both studies, we implemented an object search task which is usually included in MR remote collaboration experiments [3,41], and is common in real-world practice (e.g. an expert guides a trainee to find tools for industrial maintenance in a dynamic and cluttered environment). With our MR system, a local worker wearing an AR display is guided to target locations in the real world by a remote expert in a VR interface. The remote expert is able to virtually move around and place virtual auditory beacons in the local worker's physical environment.

As illustrated in Fig. 1, the remote expert's voice and the auditory beacons are spatialized in the local space. When the visual cues are enabled, virtual representations of the remote expert's head frustum and hand gesture are also visible to the local worker through the AR headset. The local worker's egocentric view is live captured by the AR HMD and shared for the remote expert to see the real environment. Our work shows that a local worker can be navigated to small (2 cm^3) real Lego bricks in a large office (90 m^2) with strong spatial perception by only using spatial auditory cues. In addition to the spatial auditory cues, integrating visual cues, especially the remote expert's head frustum, can further enhance local workers' task performance as well as their social presence and spatial awareness.

Compared to prior work, the main contributions of this paper include:

1. A MR remote collaboration system that spatializes the remote user's voice and auditory beacons, as well as two studies using the system in a large office sized workspace.
2. One of the first user studies that explores how spatial auditory cues can help a MR remote collaboration task and influence the local worker's experience.
3. One of the first user studies that explores how hybrid spatial auditory and visual cues can help a MR remote collaboration task and influence the local worker's experience.
4. An analysis of the verbal communication pattern between the local and remote users in a MR remote collaboration task under different conditions.

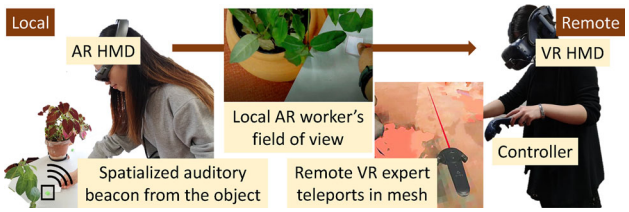

Fig. 1 The overview of our MR remote collaboration system: the remote expert uses VR teleportation to virtually move in a 3D virtual replica of the local space. The expert also talks to the local worker and provides spatialized auditory beacons that are perceived via the local worker's AR headset. When visual cues are enabled, the expert's hand and head representations will be visible to the local worker. The local worker's egocentric view is always live shared with the remote expert

After a short review of related works, we present the technical details of our system, describe our two user studies and findings, and discuss the results and design implications for future MR remote collaboration systems.

2 Related work

Remote collaboration has become more and more convenient with the development of commercial multi-party video-conference systems that are commonly used nowadays. The demand for conveying more information and enhancing the sense of presence has driven researchers to explore the sharing of face-to-face non-verbal communication cues. This has been possible thanks to infrastructure and AR/VR hardware advancements, which evolved traditional remote collaboration interfaces to spatial telepresence systems, such as Holoportation [32].

2.1 Sharing scenes from local to remote

In MR remote collaboration, the local worker's surroundings are usually shared as video feeds to the remote expert, who can then see the captured real environment or even virtually immerse him/herself in the scene in real time. Some researchers have used head-mounted cameras [26] to capture and share the local user's first-person view, but the view angle is often very restricted, and 2D video cannot deliver stereoscopic depth information. To overcome these limitations, some researchers have explored 3D or 360° scene capturing and sharing [1,28,29,41,42]. They found that remote users could perceive a space better and move around in the scene

independently of the local worker's view with 3D or 360° video. However, 3D video sharing typically covers a small local workspace, or the scene update rate is low due to a significant increase in bandwidth requirements.

In our case, the remote expert needs to virtually move around, talk, and enable auditory beacons at different locations in a captured 3D scene of a large local area. The expert also needs to see the local worker's view from a 2D video feed for guiding the task. Therefore, similar to previous work that combined a 3D static model with a 2D live video feed [16], we import a pre-modeled 3D mesh of the real space in the VR environment for the remote expert to move around using VR teleportation. We also live stream 2D video feed(s) from the local worker's AR headset to the remote expert's headset.

2.2 Auditory cues in remote collaboration

The basic auditory medium in remote collaboration is users' speech. Some researchers have spatialized the remote users' voice in the local user's space based on the locations where the remote users' images were rendered [6,7,14]. This improved the local user's sense of engagement in a collaborative experience. Researchers have also spatialized sounds from objects in a shared scene [18] to enhance the local user's spatial perception in a telepresence system.

However, auditory cues can do more than merely enhancing an immersive experience. Research in audio AR/VR has shown that spatialized auditory beacons can aid navigation [8,30] or help localize a target in an environment [4,37,39,50,51]. This inspired us to explore how spatialized auditory cues could be used in MR remote collaboration to aid with tasks such as searching for and manipulating objects [3,41]. Therefore, we developed a MR system that spatializes the remote user's voice as well as the auditory beacons that are virtually attached to objects in various places.

2.3 Visual cues in remote collaboration

A large number of studies have shown that visual cues are essential to improve users' performance and enhance their social experience in MR remote collaboration [15,29,33,41,42]. For example, a virtual annotation is helpful in indicating the user's attention. Researchers have used devices, such as VR controllers or a mouse, to draw augmented annotations or cursor pointers [13,17,25,45]. While this is easy and precise to operate, some researchers have found that hand gestures are good indicators of rotations and orientations [2], and can be intuitively perceived by collaborators [40]. Many research projects used hand gestures to point at objects, demonstrate operations, and give directions [5,23,38]. Another common visual cue is eye gaze [19,27]. However, gaze may be difficult to interpret as an explicit location cue as it may move around due to subconscious and subtle eye movements [31]. Virtual avatars are also widely used as a visual cue to represent the partner's location in each user's own space, enhancing the users' social presence and improving their task performance [24,33].

We found that most academic and commercial research on MR remote collaboration has focused mainly on visual cues. In contrast, our system can transmit both spatial auditory and visual cues in remote communication. In addition to exploring the effectiveness of the auditory cues, we also investigated the reciprocity between these two cues and how their combination can improve the local-remote collaboration. We conducted studies in an unmodified office in regular working hours rather than in a controlled environment since we aimed to mimic potential real-life collaborative scenarios.

3 System overview

Our prototype system transmits both spatial auditory and visual cues from a remote expert to a local worker to help them collaborate on a real-world search task. The local worker wears an optical see-through AR display (Magic Leap One) that has a set of depth-sensing units for mapping out the worker's real workspace. The remote expert wears a VR headset (VIVE Pro Eye) with an externally mounted gesture sensor (Leap Motion). In the following, we elaborate on three major parts of the system and the implementation details.

3.1 Local-remote position coordination

In our studies, a local worker in AR collaborates with a remote expert in VR. Given this setup, the local AR scene and the remote VR scene need to be coordinated to share the cues at the correct positions for each other. We first create a 1:1 3D mesh of the physical space using the Magic Leap One application Dotty Mesh.[1] This 3D mesh includes the local room, the remote room, and the area connecting these two places. We import the mesh into the VR scene in which the remote expert can move around using standard VR teleportation. We align the AR and VR scenes to the same shared virtual coordinate system, in which the origin is located at the center of an image marker that is placed in the remote expert's room. On the AR side, we use the image tracking function of Magic Leap One to detect the image marker in the physical world. On the VR side, the marker center in the virtual mesh is situated exactly at its corresponding location in the physical world. In this way, both the AR and VR users' movements can be projected with respect to the same virtual origin and then placed at the correct positions in the partner's own space, without further position or pose transformation. With this local-remote position calibration, both

[1] https://dottyar.com/.

users can feel co-located in the same space, similar to face-to-face interactions.

3.2 Sharing of auditory cues

Auditory information can provide an overall perception of the environment. In the study, we integrate three types of auditory cues in our system:

1. Non-spatial voice: The local worker and the remote expert can talk like in an audio call with non-spatial voice. This is enabled by connecting both AR and VR sides to the same network for exchanging voice data using the Dissonance Unity Voice Chat plugin.[2]
2. Spatialized voice: The local worker talks with non-spatial voice as above, but the remote expert's voice is spatialized to the local worker, i.e. the worker can hear the expert's voice from a location in the real world that corresponds to the position in the modeled VR scene where the expert has virtually placed him/herself by VR teleportation.
3. Spatialized auditory beacons: The remote expert can virtually attach and play auditory beacons at target objects by using the HTC Vive controller to interact with the object representations in the VR scene. The auditory beacon is spatialized from the object to the local worker. The beacon is a 2 s long but a looping, wide-band musical sound designed by the authors. Its peak corresponds with the human ear's elevated frequency response (2.5–5 KHz).

Audio spatialization is implemented using the Magic Leap Soundfield Audio Plugin for Unity.[3] To impart some realism to the virtually spatialized auditory cues 2 and 3, we rendered real-life sound properties such as the changes in the frequency spectrum and sound level based on the user's orientation and distance from the sound source. We can also spatialize the local worker's voice to the remote side, but we omit this in the current prototype because we focus on investigating the local worker's experience.

3.3 Sharing of visual cues

We integrate visual cues to explore their combination with auditory cues in MR remote collaboration. Considering the advantages and disadvantages discussed in Sect. 2, we share two visual cues from the remote to the local side:

1. Head frustum: The head frustum represents the remote expert's head position, viewing direction, and FoV. As shown in Fig. 2a, we use a red sphere for the head, a black cuboid for the eyes, and a yellow frustum for the viewing direction that extends out from the cuboid (eyes). The head frustum affords large-scale spatial guidance. It is always visible to the local worker, reflecting the remote expert's movements in the local environment.
2. Hand gesture: As shown in Fig. 2b, a virtual 3D mesh of the remote expert's hand gesture is captured by the Leap Motion and overlaid onto the local worker's view. The hand gesture is for more precise target indication in the search task. This can be especially useful if the user cannot locate the target among several adjacent ones following the head frustum and/or the auditory beacons. The hand gesture is only visible when the remote expert proactively uses hand(s) to indicate the target object in a small area like a desk.

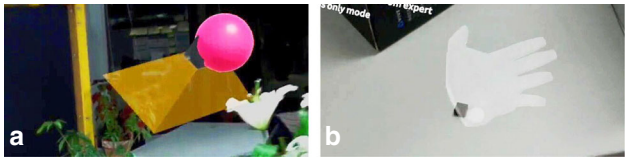

Fig. 2 The illustrations of head frustum (**a**) and hand gesture (**b**) shared from the remote to the local side

From the local to the remote side, we also visualize the local worker's head frustum in the VR headset to illustrate the worker's movements. The local worker's FoV is captured by the Magic Leap One's built-in camera and live streamed for the remote expert to see the physical environment from the local worker's viewpoint.

3.4 Implementation

Our prototype system was built with VR and AR HMDs for the remote expert and the local worker respectively. The remote expert used the VIVE Pro Eye that was connected to a desktop computer (Intel Core i7-8700 3.2 GHz CPU with 6 Cores, 32 GB DDR4 RAM, NVIDIA GeForce GTX 1080 GPU) running the Windows 10 OS. A Leap Motion hand gesture sensor was mounted on the front panel of the VR headset. The local worker used a Magic Leap One AR display to perceive the spatialized auditory and visual cues. The Magic Leap One camera was first used to track the image marker for the local-remote coordination, and then used to capture and live stream the local worker's FoV. The camera is located next to the right eye, so the captured FoV was marginally offset from the user's actual FoV. However, this had little influence on our study, since the remote expert only needed this information as a rough guide of the local worker's movements and surroundings.

[2] https://placeholder-software.co.uk/dissonance/.

[3] https://creator.magicleap.com/learn/guides/soundfield-user-guide-for-unity.

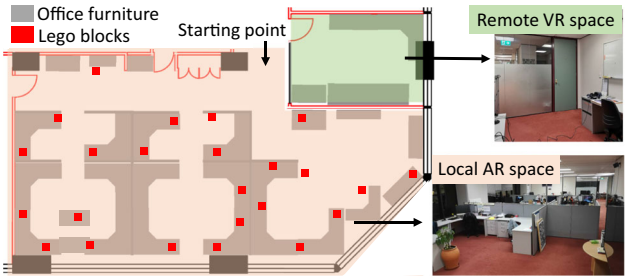

Fig. 3 The user study area: the space layout and the locations of 24 Lego bricks in Study 1 are shown in the floor plan at the left. The pictures of the local and the remote spaces are shown at the right. The Lego bricks were placed at different heights and were rearranged in Study 2

The system was developed using the Unity3D game engine (2019.1.7f1). Both AR and VR sides were connected to the same Wifi network for position synchronization and audio/visual data exchange. The local worker used the integrated microphone of the Magic Leap to communicate with the remote expert. The auditory cues were delivered over the speakers of Magic Leap. The remote expert heard the local user using the integrated headphones of Vive Pro, and used an external microphone for voice communication.

4 User study 1: Spatial auditory cues

In the first user study, we investigated how different auditory cues (voice and auditory beacon) could influence the participants' performance and experience with our system in MR remote collaboration tasks. We conducted an object search task, in which the remote expert used auditory cues to guide the local worker to find real Lego bricks in the local space. Study 1 (S1) was motivated by the following research questions (RQ):

- S1-RQ1. How efficiently can spatial auditory cues help a local worker navigate an environment in a MR remote collaborative search task?
- S1-RQ2. Can spatial auditory cues enhance a local worker's perception of the spatial layout and the remote partner's co-presence?
- S1-RQ3. How do different auditory cue conditions influence the local-remote conversation pattern?

4.1 Study environment and setup

As shown in Fig. 3, two adjacent but physically isolated spaces on the same floor were used for the user study. The local worker conducted object search tasks in an unmodified office of approximately 90 m^2 with 15 desks, a small empty area, and various furniture and office supplies. The remote expert was in a separate room of around 15 m^2. The physical separation between the remote expert and the local worker made their conversation only possible via the MR collaboration system. The studies were carried out during regular office hours without isolating participants from the office environment. We believe this natural environment setup lends some level of ecological validity to the study by mimicking potential applications in dynamic and cluttered situations.

The prototype system as described in Sect. 3 was used for both studies. Participants were recruited as the local AR workers, while only one trained VR user (female, 27) played the role of the remote expert in both studies. This was because our research focus was on the local side, and using the VR interface required some training, so the remote expert's familiarity with the system could influence the local worker's performance and experience. Therefore, having the same remote expert could keep the communication consistent, and reduce unwanted effects that may arise from unversed or incorrect operations from the remote side.

The same environment was also used in Study 2 that explored the combination of auditory and visual cues.

4.2 Experimental design

The experiment was designed as a within-subjects study using the following four auditory conditions A1-A4:

- A1. Non-spatialized voice: Both users talked to each other like in a normal phone or video call.
- A2. Spatialized voice: The remote expert's voice was spatialized in the local space, but the local worker's voice was not spatialized to the remote expert.
- A3. Non-spatialized voice + spatialized auditory beacon: Both users' voice was delivered as in A1, but an auditory beacon was binaurally spatialized from the target object to the local user, i.e. virtual sound was placed onto the target object.
- A4. Spatialized voice + spatialized auditory beacon: Users' voice was delivered as in A2, and the auditory beacon was binaurally spatialized as in A3.

4.3 Experimental task

The target objects of our search task were 24 Lego bricks of the same shape and size (1.5 cm × 1.5 cm × 0.9 cm). The bricks were placed at different heights in the local space, and some were located in areas of high clutter and/or near objects with similar colors, which partially occluded the bricks or introduced visual distractions, as shown in Fig. 4.

For each cue condition, i.e. one trial, the expert guided a participant to find four Lego bricks. The participant started each trial at the same point as marked in Fig. 3 and returned

Fig. 4 Examples of the Lego brick layout in Study 1. We placed the bricks with occlusions and/or close to objects with similar colors

there after finding all four bricks. Upon finding each brick, the participant would leave a small paper tag next to it, and then look for the next one following the cue. In each trial, the remote expert arbitrarily selected four target bricks with the following considerations: (1) the selection of bricks was basically counterbalanced regarding the physical distance a participant was supposed to walk, lending some validity to the evaluation of participants' performance; (2) the selection of bricks was basically counterbalanced regarding the search difficulty that was influenced by visual occlusions, color similarity of the bricks to their surroundings, etc.

In A1 and A2, only verbal instructions were available. Explicit guidance such as "*go to the printer*" and "*turn left*" was given to the participant. A2 provided some spatial indications by spatializing the remote expert's voice. As for A3 and A4, spatialized auditory beacons were added along with verbal communication. To study the beacons' efficacy more fairly, the expert assumed a more observational role in A3 and A4. The expert barely gave explicit guidance but maintained a smooth and interactive collaboration with the participants using some simple communication, which included answering questions (e.g. "*yes, this is the correct direction*") and the transition between bricks along with the beacon activation (e.g. "*okay, then the next object*…"). The remote expert only activated one beacon each time, muted it once the target was found, and then activated the next one.

In all four conditions, the expert verbally (1) answered participants' questions and (2) confirmed each target finding and transited to the next search. Every participant talked spontaneously and differently, so the expert's communication could not be exactly the same all the time, but the expert's guidance procedure stayed consistent with similar wording, e.g. "*the first object is...yes, you are correct, then the next one is...okay, you have found all, you can come back*".

During the entire search task, the local participant's view was live streamed to the remote side and shown in the VR scene together with the 3D mesh of the local space. The remote expert moved through the mesh using the standard VR teleportation technique (i.e. moving to desired locations in the mesh using the VR headset controller). Via the teleportation, the remote expert guided the search task by "walking" around and moving towards the selected Lego bricks to enable the auditory beacons in the virtual scene. Since the expert's virtual movements were projected back to the AR coordinate in the local environment, the auditory beacon and the expert's voice were binaurally spatialized from corresponding locations in the local space in real time.

4.4 Measurements

For objective measures, we recorded the task completion time in a system log file to measure the participants' performance in each cue condition. For subjective feedback, we asked participants to fill in questionnaires after each cue condition. We used the Networked Mind Measure of Social Presence Questionnaire (SoPQ) [21] for measuring social presence, the NASA Task Load Index Questionnaire (NASA-TLX) [34] for measuring workload, and the System Usability Scale (SUS) [11] for measuring the usability of the system.

We also asked participants to fill in the MEC Spatial Presence Questionnaire (SpPQ) [47]. SpPQ is usually used to measure the remote user's spatial awareness in the virtual environment, but we adopted this questionnaire on the local side to assess how the virtually spatialized auditory cues affected the local workers' spatial perception in the physical world. To this end, we only used the fitting subscale of this questionnaire, Spatial Situation Model (SSM). As shown in Table 1, the original SSM questions barely fit our context, so we slightly adjusted these questions while keeping their original intentions. For example, the original Q3 and Q4 ask about two typical aspects in understanding a space: *size* and *distance*. Since *size* was not applicable to sound sources, we asked direction and distance instead, which corresponded to the original questions by also asking a user's spatial perception, but specifically with our provided cues.

After all four conditions, participants filled in a post-experiment questionnaire for overall condition ranking and further comments. We audio-recorded the user study in order to analyze the conversation patterns in each cue condition.

4.5 Procedure

We started the experiment by introducing the study to the participant. Then, the participant signed a consent form and answered demographic questions. After that, the participant put on the audio recorder and the Magic Leap One, aligned the AR and VR coordinates in the remote room, then went out and stood at the starting point. When both users were ready, the first trial started with a randomly selected cue condition. After finding four bricks following the auditory beacons and/or the remote expert's verbal communication, the participant went back to the remote room and filled in the questionnaires. After that, the participant put on the devices again and repeated the above process for the other three conditions. In the end, the participant filled in a post-experiment questionnaire to compare and comment about all the cue conditions. We finished the study with a short informal interview

Table 1 The original questions, adjusted questions in Study 1, and adjusted questions in Study 2 of the SSM questionnaire

	Original question	Customized question in Study 1	Customized question in Study 2
Q1	I was able to imagine the arrangement of the spaces in the medium very well	I was able to imagine the spatial locations of the speaking person and/or the sounding objects in the medium very well	I was able to imagine the spatial locations of the local objects and the remote person in the medium very well
Q2	I had a precise idea of the spatial surroundings presented in the medium	I had a precise idea of the spatial locations of the speaking person and/or the sounding objects presented in the medium	I had a precise idea of the spatial locations of the local objects and the remote person presented in the medium
Q3	I was able to make a good estimate of the size of the presented space	I was able to make a good estimate of the distance between the presented voice (and sound) and myself	I was able to make a good estimate of the distance between the remote person and myself, as well as the distance between the local objects and myself
Q4	I was able to make a good estimate of how far apart things were from each other	I was able to make a good estimate of the direction of the presented voice (and sound) from myself	I was able to make a good estimate of the direction of the remote person from myself, as well as the direction of the local objects from myself
Q5	Even now, I still have a concrete mental image of the spatial environment	Even now, I still have a concrete mental image of the spatial soundscape of the environment	Even now, I still have a concrete mental image of the spatial layout of the local objects in the environment
Q6	Even now, I could still find my way around the spatial environment in the presentation	Even now, I could still find my way to the spatial sound sources in the presentation	Even now, I could still find my way to the local objects in the presentation

with the participant. The order of conditions was counterbalanced among participants and each condition was run with four arbitrarily selected Lego bricks.

5 Results and discussion of Study 1

In this section, we report on the results of Study 1, statistical analysis ($\alpha = .05$), and effect sizes (ES). While the statistical significance shows the *probability* of an observed difference being due to chance, the ES implies the *magnitude* of such a difference [44]. The ES $\in [0, 1]$ interprets 0.1 as small effect, 0.3 as moderate effect, and above 0.5 as strong effect.[4,5] We also answer S1-RQs by discussing the study results, our observations, and participants' feedback.

A total of 24 participants (15 male, 9 female, age $\in [20, 40]$, mean = 26.4, SD = 4.271) with normal hearing and vision took part in the study. Only five participants reported familiarity with AR interfaces as they were using AR systems a few times a month. The others had either limited exposure to AR devices or had never tried it before. None of the participants were familiar with the experiment environment.

[4] https://www.sheffield.ac.uk/polopoly_fs/1.714575!/file/stcp-marshall-FriedmanS.pdf.

[5] https://www.sheffield.ac.uk/polopoly_fs/1.714573!/file/stcp-marshall-WilcoxonS.pdf.

5.1 Results of the measurements

5.1.1 Task completion time

On average, participants spent similar amounts of time on each condition (A1: 118.88 s, A2: 122.75 s, A3: 116 s, A4: 120.21 s). Some conditions did not follow normal distribution based on a Shapiro-Wilk test, so we ran a Friedman test and found no significant difference ($\chi^2(3) = 3.567$, $p = .312$). A follow-up Kendall's W test showed a small difference magnitude between different conditions ($ES = .05$). The result indicates no significant difference in task completion time between all cue conditions.

5.1.2 Social presence

To investigate if different cues affected the participants' experience with presence and attention, we used the Social Presence questionnaire (SoPQ) [21] including three subscales: Co-Presence (CP), Attention Allocation (AA), and Perceived Message Understanding (PMU). The whole questionnaire has 18 rating items on a 7-point Likert scale (1: strongly disagree–7: strongly agree). A Friedman test and a Kendall's W test showed that all subscales had no significant difference with a small ES between cue conditions: CP ($\chi^2(3) = 5.15$, $p = .161$, $ES = .012$), AA ($\chi^2(3) =$

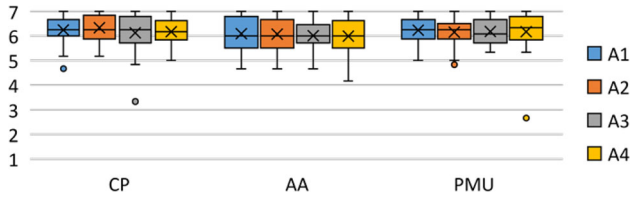

Fig. 5 Results of the Social Presence questionnaire. *CP*, Co-Presence; *AA*, Attention Allocation; *PMU*, Perceived Message Understanding

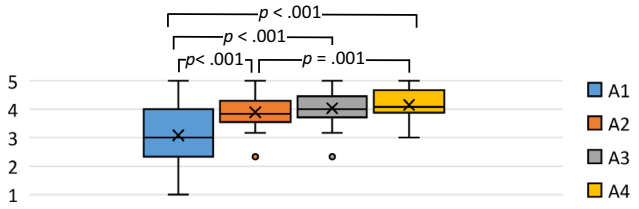

Fig. 6 Results of the Spatial Situation Model questionnaire

6.239, $p = .101$, $ES = .014$), PMU ($\chi^2(3) = 3.162$, $p = .367$, $ES = .007$). However, as shown in Fig. 5, we see a generally high rating of all conditions on each subscale (> 5.95 out of 7). This indicates that the participants generally had a strong social presence experience with the provided auditory cues.

5.1.3 Spatial presence

The subscale SSM of the SpPQ questionnaire [47] was used to investigate the participants' spatial awareness with different auditory cues. It consists of six rating items on a 5-point Likert scale (1: fully disagree–5: fully agree). The results are shown in Fig. 6. A Friedman test indicated significant difference across the cues ($\chi^2(3) = 106.097$, $p < .001$, $ES = .246$). A post hoc analysis with the Wilcoxon signed-rank test showed significant pairwise differences except A2–A3 ($Z = -1.573$, $p = .116$, $ES = .131$) and A3–A4 ($Z = -1.467$, $p = .142$, $ES = .122$). The ES of Wilcoxon signed-rank test is calculated by $Z/sqrt(N)$, in which N is the amount of participants. The significantly different cue pairs also showed moderate to large ES $\in$ [.284, .665]. The results suggest that participants had a significantly improved spatial perception of the Lego bricks and the remote expert with spatialized auditory cues, but the difference between these spatial auditory cues was not strong (except A2–A4).

5.1.4 Workload

We used the NASA-TLX questionnaire [34] to compare the participants' physical and mental workload across conditions. NASA-TLX includes six rating items in a 100-point range with 5-point steps (0: very low–100: very high, the lower, the better). We focused on three most relevant items in our study: mental demand, effort, and frustration. A Friedman test and a Kendall's W test showed no significant difference and very small ES of cue conditions on the participants' workload (mental demand: $\chi^2(3) = 2.986$, $p = .394$, $ES = .041$, effort: $\chi^2(3) = 1.667$, $p = .644$, $ES = .023$, frustration: $\chi^2(3) = 1.703$, $p = .636$, $ES = .024$). As shown in Fig. 7, most participants did not experience much frustration, but it required some mental demand and effort to complete the task, no matter which auditory condition was used.

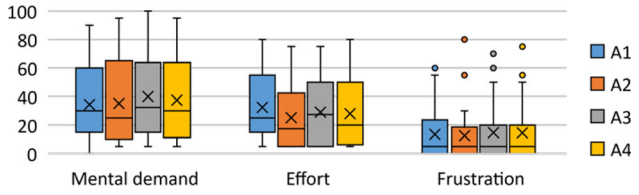

Fig. 7 Results of the NASA-TLX questionnaire. We focus on three most relevant rating items: mental demand, effort, and frustration

5.1.5 System usability

We used the SUS [11] to evaluate how the participants would assess the usability of our system. The results are summarized in Table 2. A Friedman test showed no significant difference ($\chi^2(3) = .237$, $p = .971$, $ES = .003$) between conditions on the participants' usability assessment. However, participants overall felt our system had above-average usability since the SUS scores were above 68 [10].

5.1.6 User preference

We show the participants' preference ranking in Fig. 8 for this remote collaboration task. For each condition, the ranking results were represented with numbers 1 (Rank1)–4 (Rank4) before running the significance tests. We found significant difference with moderate ES across conditions ($\chi^2(3) = 29.75$, $p < .001$, $ES = .413$) through a Friedman test and a Kendall's W test. We then ran Wilcoxon signed-rank tests to investigate pairwise differences and found that except A1-A2 ($Z = -.323$, $p = .747$, $ES = .066$) and A3-A4 ($Z = -1.060$, $p = .289$, $ES = .216$), all the other pairs had significantly different influence on the ranking with large ES $\in$ [.609, .779]. The result shows that most participants strongly preferred to use A3 and A4, i.e. non-spatialized/spatialized voice + spatialized auditory beacon.

A total of 11 participants commented that the spatialized auditory beacon provided clear direction and distance to follow, e.g. "*Spatial sound from objects was easy to perceive in terms of distance and direction to object*" (P7, male, 29). It is a bit surprising that the non-spatialized voice was slightly preferred over spatialized voice. This could be because our audio spatialization involved the effect of distance fall-off.

Table 2 The SUS mean score, median score, and SD value for each auditory condition

	A1	A2	A3	A4
Mean	76.56	76.98	77.92	78.13
Median	76.25	76.25	75	76.25
SD	11.768	10.268	15.702	14.128

SUS score $\in$ [0, 100], the higher, the better

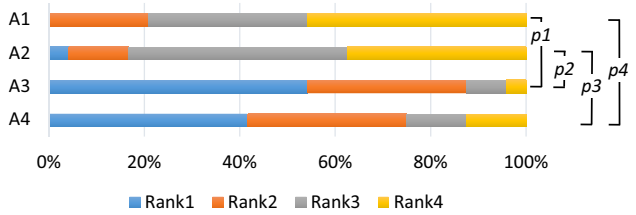

Fig. 8 User preference ranking of four auditory cues. Rank1 is the most preferred. $p1 < .001$, $p2 < .001$, $p3 = .003$, $p4 = .002$

This was more real, but weakened the expert's voice when being far away, which affected some participants' perception.

5.2 Discussion of the measurement results

In the following, we discuss the measurement results, participants' feedback, possible reasons for some results, and implications from the study. We will answer S1-RQ1 and S1-RQ2. S1-RQ3 about conversation pattern will be detailed in the next subsection.

S1-RQ1 is about the efficiency of the spatial auditory cues in navigating the environment for the search task. The results of task completion time did not show significant difference between non-spatialized and spatialized auditory cues. However, considering that the participants could rely on explicit verbal instructions in A1 and A2 but mainly used auditory beacons in A3 and A4 to search bricks, the results might indicate insignificant difference between detailed verbal description and spatialized auditory beacons for the search task in this environment. Some participants commented that "*the audio cue was useful and clearly showed me the direction*" (P7, male, 29) and "*very intuitive to follow the spatial sound*" (P14, male, 25). Section 5.1.6 mentioned a distance fall-off effect in the audio spatialization. This effect did not cause problems with perceiving the auditory beacons, which could be because the default volume of the beacon was set properly. When the expert's spatialized voice was clearly perceived, some participants could "*follow the direction of the voice and move to the correct area*" (P5, female, 24).

S1-RQ2 is about the participants' spatial awareness and co-presence experience. Regarding their spatial awareness, it is unsurprising that the audio spatialization significantly enhanced participants' spatial perception of the Lego bricks and of the remote partner, as shown in Fig. 6. As for the co-presence experience, there was no significant difference between cue conditions as shown in Fig. 5, but all conditions were rated high (average above 6), which indicates that all conditions produced a good co-presence experience for some reason. The positive aspects of spatialized audio could be the spatial effects like in a real-world scenario. When the expert simply communicated the transition between the bricks (e.g. "*yes, you are correct, then the next one is…*") when virtually moving around, the spatial effects helped some participants "*feel the presence of my partner*" (P2, male, 30).

It is surprising that the non-spatialized verbal communication (A1), like a phone call, also produced good social presence. This could be partly because "*the instructions are very descriptive*" (P23, female, 22). Plus, the expert instantly answered participants' questions and confirmed or corrected their movements based on the live streamed participants' FoV, which made an impression of "*smooth and natural communication*" (P10, female, 23), like the expert was present in the environment. These findings might indicate the importance and/or the influence of verbal communication in MR remote collaboration, which was possibly overlooked or not deeply analysed in existing work that focused on visual cues but also included verbal communication [3,41]. However, participants' positive social presence with A1 could also benefit from the remote expert's familiarity with the system and with the experimental environment. In addition, when only enabling verbal communication, the remote expert had a heavy workload to give detailed instructions.

We have discussed some advantages of explicit verbal communication (e.g. straightforward to understand) and spatialized auditory cues (e.g. intuitive to follow). They also had some drawbacks. With only verbal instructions, participants had to "*listen very carefully*" (P21, female, 24) and did not have other information to actively direct themselves. Spatialized audio "*could be distracting*" (P3, male, 27) with the volume change and might confuse people who did not listen clearly. Overall, participants acknowledged the usability of our system with all cue conditions, but when ranking the conditions, most people preferred spatialized auditory cues, especially the auditory beacons (A3 and A4), over only verbal instructions.

During the study, the participants' familiarity with the environment inevitably increased as the trials went on, but its influence should have been limited for the following reasons. First, we counterbalanced the conditions and each condition was run with four arbitrary bricks. Moreover, although several bricks were put close to each other on purpose, most bricks were well distributed and/or well hidden in the space, so it was unlikely that participants noticed most of them during the first few conditions.

The results of Study 1 show us the effects of different auditory cues in a collaborative search task. The study provides insights into the influence of the verbal communication

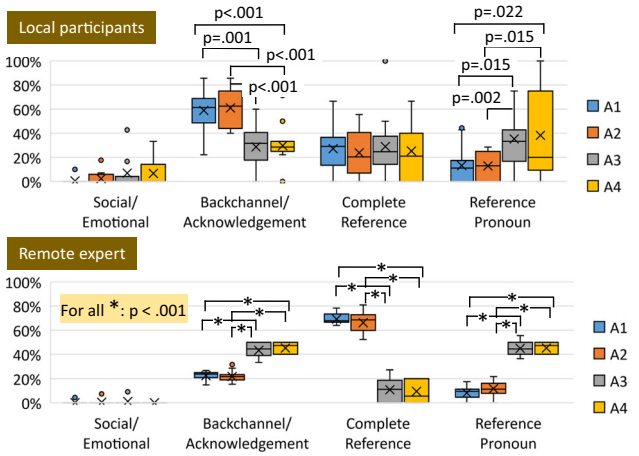

Fig. 9 Study 1: percentage of the utterance types in each participant's conversation with the remote expert. As some conditions did not follow normal distribution, we used Friedman tests and then Wilcoxon signed-rank tests to check significance between conditions

on the task performance and social presence experience. The results also encourage us to use spatialized auditory cues for local workers to intuitively experience the surroundings and the remote user when conducting a task. In a real-world scenario, a real user might not use an auditory beacon to guide a search task. However, in a remote collaboration supported by MR techniques, the spatialized auditory beacons might be used as an additional channel of information.

5.3 Conversation pattern and non-verbal behavior

In the following, we discuss the local-remote conversation pattern and the participants' non-verbal behavior during the task. We will answer S1-RQ3 and discuss how the conditions influenced the participants' interaction with the remote expert. We analyzed 22 participants' conversations and two participants were excluded due to recording failure.

According to the work by Smith and Neff [36], we categorized verbal communications into four types of utterance: social/emotional, backchannel/acknowledgement, complete reference, and reference pronoun. Utterance refers to a section of speech (a sentence or comparable). *Social/emotional* means expressions of feelings like "wow this is great". *Backchannel/ acknowledgement* means indications of listening like "okay". *Complete reference* means utterances that can be understood by itself like "behind the yellow table". *Reference pronoun* uses pronouns to refer to things like "behind *this* table". We count the remote expert's confirmation of each brick finding as *backchannel/acknowledgement*, and the transition to the next search in A3 and A4 ("the next object is *this* one") as *reference pronoun*. In Fig. 9, we plot the percentage of each utterance type in each cue condition for both local and remote users. In the figure we also mark the pairwise significance values as per Wilcoxon signed-rank tests.

On the remote side, when the expert gave explicit instructions in A1 and A2, it is unsurprising that most utterances were *complete reference*. When auditory beacons were included (A3 and A4), the percentage of *complete reference* significantly decreased, and most utterances were *backchannel/acknowledgement* (for brick confirmation) and *reference pronoun* (for initializing the next search). In some cases, the expert still needed to give hints (*complete* or *pronoun* references) when participants could not follow the auditory beacon correctly or when they asked for clarification.

More interesting findings were discovered on the local side. In A1 and A2, we can see that a large portion of the participants' utterances was *backchannel/acknowledgement*. This was because participants usually only responded a simple "*yes*" or "*okay*" to the expert's instructions. Some participants repeated the expert's descriptions using *complete reference* like in a thinking-aloud process, e.g. "*hmm, the left edge of the yellow desk*". This shows that participants mainly followed the expert, instead of being in an active interaction. The conversation pattern of A2 was similar to that of A1 but we noticed that a few participants were willing to check the expert's position by asking, e.g. "*where are you actually?*".

As for the non-verbal behavior, participants' movements were generally passive in A1 and A2. They normally waited for the expert's instructions before they moved or adjusted their direction. As the environment was new to them, some participants first looked around when they got instructions with a specific landmark. In A2, a few participants could follow the direction of the expert's voice before getting a complete instruction.

When auditory beacons were enabled (A3 and A4), we observed that some participants immediately recognized the orientation of the sound source and quickly approached it, but some participants first turned their head a bit to confirm the orientation before moving towards the sound source. For bricks that were not far from each other, some participants recognized the correct one without much hesitation, but some carefully approached and listened to each of them before determining the target. Besides individual listening sensitivity, this difference could also be partly because some participants' initial orientation when approaching the bricks enabled them to recognize the source easily.

The participants' active movement corresponds with the significantly reduced *backchannel/acknowledgement* utterances in A3 and A4. Some participants actively checked their direction with the remote expert using *complete* or *pronoun* references, e.g. "*am I in the correct direction?*", "*I think it is around here?*". This could be because the participants would like some connection with their partner, or they were not completely confident about their spatial perception. Although the portion of *social/emotional* utterances did not significantly increase in A3 and A4, we noticed some participants' expression of their feeling, e.g. "*wow it is cool*". These changes

indicate some interactions with the remote expert, instead of just following the instructions.

6 User study 2: Spatial auditory + visual cues

After Study 1, we conducted a second study, Study 2 (S2), to explore how the hybrid spatial auditory + visual cues could affect the local workers' performance and experience. We used the same object search task to investigate the following research questions:

S2-RQ1. How much can visual cues improve the participants' performance and experience compared with the audio-only condition?
S2-RQ2. Will visual cues outperform auditory cues? i.e. Will participants ignore the spatial auditory cues and mainly follow the visual cues to search bricks?
S2-RQ3. Is there any change in the local-remote conversation pattern when visual cues are integrated?

This study was also designed as a within-subjects study using the following four hybrid conditions H1–H4:

H1. Spatialized voice + spatialized auditory beacon: This condition, same as A4 in Study 1, served as the control condition in Study 2. Study 1 has shown that spatialized audio significantly improved participants' spatial awareness when guiding the search task. We also assumed that this condition would have better compatibility with the visual cues than the other auditory cues. For example, although A3 (with non-spatialized voice) showed comparable results, we assumed that it would be more natural to hear a spatialized voice while seeing a virtual head frustum moving around.
H2. H1 + hand gestures: The remote expert would use hand gestures to precisely indicate the target brick when it entered the local worker's FoV. Other objects and bricks in the environment might also be in the worker's view.
H3. H1 + head frustum: The head frustum of the remote expert would stay visible in the AR display, co-located with the expert when the expert moved by VR teleportation.
H4. H1 + hand gestures + head frustum: All spatial auditory and visual cues were shared to the local worker.

H1 (same as A4) was implemented in all conditions and conducted in the same way as in Study 1 (see Sect. 4.3) to more fairly evaluate the impact of visual cues. Otherwise, it would have been difficult to attribute the study result to the integration of visual cues, or the changed implementation of auditory cues.

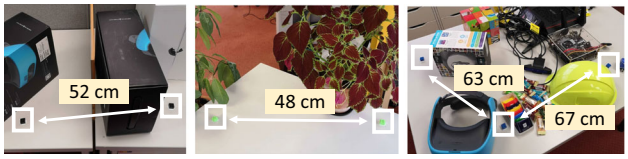

Fig. 10 Examples of the Lego brick layout in Study 2. Compared to Study 1, we increased the task difficulty by putting some bricks close to each other

In Study 2, we did not include a visual-only condition mainly due to the following considerations. As others [33,41] have shown, it is difficult to entirely exclude speech even in the presence of strong visual cues. Participants may ask questions or naturally initiate a dialogue, so excluding conversation is not advisable, but allowing dialogue may affect inter-participant consistency. Plus, our study focus is on the auditory cues. Our goal is not to present a "good cue" for a specific remote collaboration task, but to explore the appropriateness of auditory cues and their combination with visual cues. We also understand the potential limitations without a visual-only condition and will discuss the relevant issues in Sects. 7.2 and 8.

In Study 2, we used the same experimental environment, task, procedure, and measurements as in Study 1, but we applied the following changes. First, we attempted to reduce the learning effect by changing the positions and surroundings of the Lego bricks. To increase the task difficulty, we put some bricks close to each other like shown in Fig. 10. Plus, we again adjusted the original questions in the SSM questionnaire to feature the integrated visual cues. The customized questions are shown in the last column of Table 1. Other implementation details, like the counterbalance of brick selection, stayed the same as in Study 1.

7 Results and discussion of Study 2

In this section, we report on and discuss the results of Study 2. The same 24 participants took part in Study 2 one week after Study 1.

7.1 Results of the measurements

7.1.1 Task completion time

Figure 11 shows the task completion time in all cue conditions. All conditions followed normal distribution according to a Shapiro-Wilk test, so we used a repeated-measure ANOVA for factorial analysis and found an overall signif-

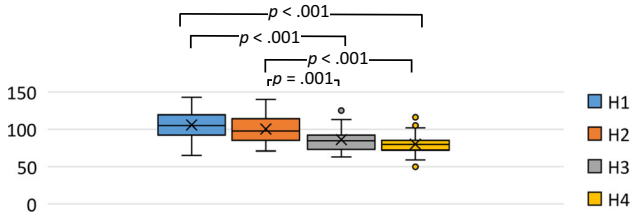

Fig. 11 Task completion time (s) in Study 2

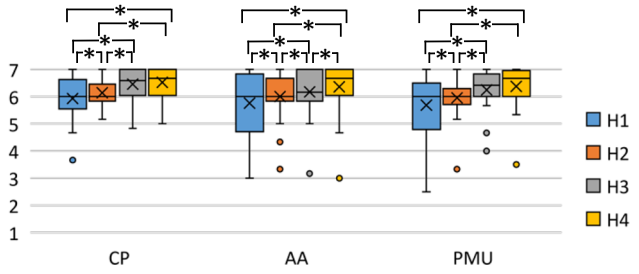

Fig. 12 Results of the Social Presence questionnaire. Most cue pairs have a significant difference (*)

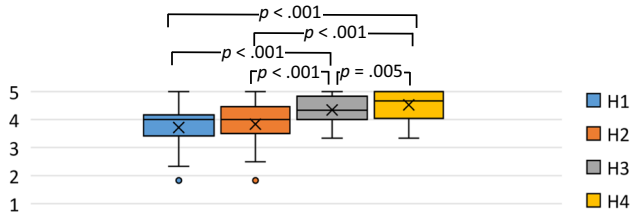

Fig. 13 Results of the Spatial Situation Model questionnaire

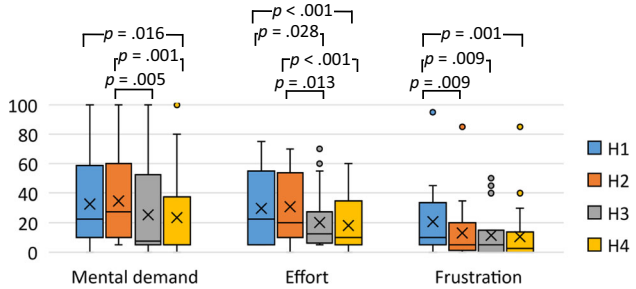

Fig. 14 Results of the NASA-TLX questionnaire

icant difference with a large effect size (ES) ($F(3, 69) = 34.092$, $p < .001$, $ES = .597$). Then, a Bonferroni post hoc test showed significant pairwise differences except H1–H2 ($p = .588$) and H3–H4 ($p = .197$). These results indicate that compared with the control condition, the integration of hand gestures did not significantly decrease the participants' task completion time, but participants finished the task significantly faster with the remote expert's head frustum.

7.1.2 Social presence

We used the subscales Co-Presence (CP), Attention Allocation (AA), and Perceived Message Understanding (PMU) to evaluate the participants' social presence experience. Friedman tests and Kendall's W tests showed significant differences with small ES across conditions for all three subscales: CP: $\chi^2(3) = 63.115$, $p < .001$, $ES = .146$, AA: $\chi^2(3) = 44.848$, $p < .001$, $ES = .104$, PMU: $\chi^2(3) = 62.127$, $p < .001$, $ES = .144$. We further used Wilcoxon signed-rank tests to examine the pairwise difference. For AA, we found significant differences in all cue pairs with an ES $\in [.163, .468]$. For CP and PMU, we found significant pairwise differences except in H3–H4 (CP: $Z = -1.094$, $p = .274$, $ES = .091$, PMU: $Z = -1.411$, $p = .158$, $ES = .118$). These results indicate that the integration of both visual cues significantly improved the participants' social presence compared with the control condition. However, when the head frustum was already provided (H3), the further integration of hand gestures (H4) did not significantly enhance CP and PMU (Fig. 12).

7.1.3 Spatial presence

Like in Study 1, we used a customized SSM questionnaire to assess how different hybrid cues affected the participants' spatial perception. A Friedman test showed significant difference across cue conditions ($\chi^2(3) = 126.799$, $p < .001$, $ES = .294$). Figure 13 shows the results with pairwise significance examined by Wilcoxon signed-rank tests. It shows that only integrating hand gestures (H2) did not significantly enhance the participants' spatial awareness compared with the control condition (H1), but the integration of head frustum gave a significantly stronger perception of the environment.

7.1.4 Workload

As before, we focused on the three most relevant items in the NASA-TLX questionnaire. As some conditions did not follow normal distribution based on a Shapiro-Wilk test, we used Friedman tests and Wilcoxon signed-rank tests to investigate the overall and the pairwise differences. As shown in Fig. 14, compared with the control condition, hand gestures and head frustum did not significantly reduce the participants' mental demand, and the hand gestures did not significantly decrease the participants' effort either. However, participants felt significantly less frustration when visual cues were added, but there was no significant difference in frustration between conditions with visual cues.

Table 3 The SUS mean score, median score, and SD value of the four hybrid cues

	H1	H2	H3	H4
Mean	73.02	72.29	79.47	80.93
Median	75	73.75	81.25	77.5
SD	16.842	16.082	14.025	17.317

7.1.5 System usability

Table 3 summarizes the participants' assessment of the system usability in conditions H1–H4. A Friedman test and a Kendall's W test showed significant difference with small ES between the conditions ($\chi^2(3) = 11.898$, $p = .008$, $ES = .165$). Then Wilcoxon signed-rank tests showed significant differences of moderate to large ES in H1–H3 ($Z = -2.002$, $p = .045$, $ES = .409$), H1–H4 ($Z = -2.397$, $p = .017$, $ES = .489$), and H2–H4 ($Z = -2.263$, $p = .024$, $ES = .462$). These results indicate that compared with the control condition, hand gestures did not make the system more usable to the participants. Plus, the usability improvement was not significant from H2 (hand gestures) to H3 (head frustum), or from H3 (head frustum) to H4 (all). It is noticeable in Table 3 that all conditions had an average usability score higher than 68, which is above the average system usability [10].

7.1.6 User preference

Figure 15 shows the participants' preference ranking of the conditions for the given task. We found a significant difference with a large ES across conditions ($\chi^2(3) = 43.650$, $p < .001$, $ES = .606$) by Friedman and Kendall's W tests. Wilcoxon signed-rank tests showed significant differences in all condition pairs with a moderate to large ES $\in [.409, .889]$. The results indicate that the participants strongly preferred to use H4 that integrated all the auditory and visual cues, followed by H3 that was without hand gestures. Head frustum was preferred over hand in this remote collaboration task. As commented by participants, the head frustum clearly indicated the expert's movements, e.g. "*I could see the head walking so I could follow it easier*" (P6, male, 27). Hand gestures were intended to precisely pinpoint the target, but some participants thought the spatialized auditory beacon was enough to distinguish close bricks that were about 50 *cm* away from each other, e.g. "*Audio + Head is good enough to clearly locate the object*" (P23, female, 22), "*hand would be useful in a case where there are many objects very close to each other, otherwise spatial sound is sufficient*" (P1, female, 30).

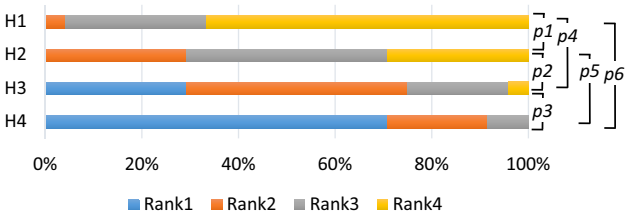

Fig. 15 User preference ranking of the four hybrid conditions. Rank1 is the most preferred. Significant differences are found in all pairs: $p1 = .025$, $p2 = .003$, $p3 = .045$, $p4 < .001$, $p5 < .001$, $p6 < .001$

7.2 Discussion of the measurement results

In the following, we discuss the above measurement results, participants' feedback, and possible reasons for some results. We will also answer S2-RQ1 and S2-RQ2.

S2-RQ1 is about the performance and experience improvement brought by the visual cues. As shown in Fig. 12, it is unsurprising that both the hand gestures and the head frustum significantly enhanced the participants' social presence compared with the auditory-only condition. This could be because both of them represent normal non-verbal cues in a face-to-face collaboration. However, according to the results of task completion time, spatial presence, effort, and system usability, head frustum significantly improved the participants' task performance and experience, but the hand gestures did not show significant effects. Compared with the head frustum that "*gave the global knowledge of where to go*" (P19, female, 24), participants had more distinct opinions about the hand gestures that were only shown in a small area for target indication. We were aware that the hand gestures could have more functionalities but we restricted them in this study, since we did not intend the hand gestures to repeat the use of the head frustum in the large-scale navigation. Plus, we assumed that hand gestures could help if the auditory beacons failed to localize closely-placed Lego bricks. However, although some participants commented that the hand gestures could "*show the specific location pretty well*" (P19), some participants felt they were redundant because "*the spatial sound is sufficient*" (P1, female, 30). This could be a reason that we did not see much difference between H1 and H2 in several metrics. Another reason could be that the small drifting of our system might have happened and such drifting influenced the perception of hand gestures more than head frustum. Head frustum was bigger than normal head size and was much more visible. However, hand gestures of the normal hand size were intended to indicate bricks precisely, thus they suffered more from position drifting. The influence of drifting was also reported in the work by Bai et al. [3], in which the issue was related to eye gaze that was represented by a raycast line.

S2-RQ2 asks if the participants would basically ignore the spatial auditory cues in the presence of visual cues. As we

did not include a visual-only condition, we could not compare how auditory cues and visual cues would function by themselves or confirm what additional values to the visual cues could be brought by spatial auditory cues. However, some participants still acknowledged the value of the spatial auditory cues in the presence of visual cues. We summarized two possible reasons based on our observations and the participants' feedback. First, the FoV of the AR HMD is very restricted but the search space was relatively large, so sometimes participants preferred to immediately acquire the direction and distance using the spatial audio. Otherwise, it was "*hard to follow when the expert moved quickly outside the FoV*" (P2, male, 30). Plus, the spatialized verbal communication might contribute to the feeling of co-presence. In this study, the remote expert talked little like in A3 and A4 of Study 1. When the expert stayed silent, some participants "*felt less connection between the partner*" (P5, female, 24) although they could still perform the task well. Some participants commented that the spatialized voice + head frustum made it more like interacting with a real person.

According to [3], we initially assumed that H4 would probably give too many cues, but participants generally commented that they did not feel a heavy workload, as also reflected in Fig. 14. This could be because the cues combined auditory and visual perception properly, and the cues were not always delivered together at the same time.

Like in Study 1, participants would inevitably become more familiar with the environment as the trials ran, especially in the sense that they had finished Study 1 in the same room the week before. However, we completely rearranged the positions of the Lego bricks and their surrounding, which would reduce the learning effect. Plus, as we discussed in Study 1, the counterbalance of the condition orders, the arbitrary selection of Lego bricks, and the distribution of the brick positions might also help alleviate the influence of learning effect on the study results.

The results of Study 2 show the improvement in the participants' task performance and experience compared to the auditory-only condition when the visual cues, especially the head frustum, were added. The combination of spatial auditory and visual cues did not induce a heavy workload to the participants. Plus, with the strong visual cues, there were still some positive comments regarding the spatial auditory cues in the sense of immediate spatial perception and a natural perception of the remote partner.

7.3 Conversation pattern and non-verbal behavior

In this study, we also analyzed the utterance percentage in the local-remote conversations as shown in Fig. 16. We will elaborate on the findings and S2-RQ3 with the discussion.

As per Friedman tests and post hoc Wilcoxon signed-rank tests, the results do not show significant changes in

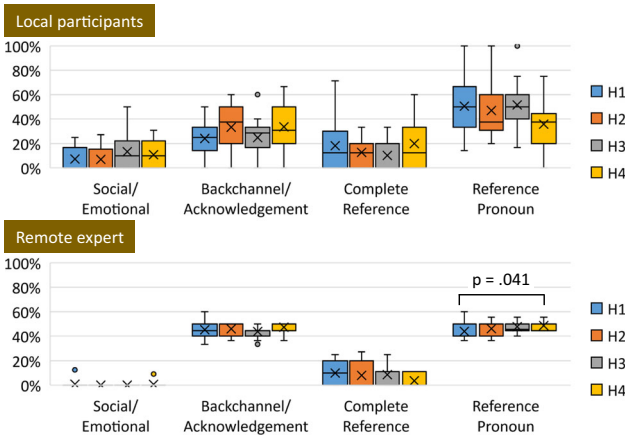

Fig. 16 Study 2: Percentage of the utterance types in each participant's conversation with the remote expert. We applied Friedman test to check significant difference (*) across conditions in each utterance type

the percentage of each utterance type between conditions, except H1–H4 in *reference pronoun* on the remote side ($Z = -2.047$, $p = .041$). However, we noticed some difference in the utterance content. In H1 that only included auditory cues, some participants still asked for clarification and confirmation using *complete* or *pronoun* references. With visual cues, the conversation became more intuitive. For example, when seeing the head frustum and the hand gestures, some participants said "*I can see you moving*" or "*ah your hand/head is here*". When visual cues were invisible in their current FoV, some participants intuitively asked "*where is your hand/head?*". The remote expert sometimes also naturally said "*do you see my head/hand*" or "*this one*" (with hand gestures) as a hint.

More interestingly, when the remote expert moved too fast, a couple of participants even asked the expert to slow down, e.g. "*it's so fast, wait for me!*". Participants also sounded more confident with visual cues. In Study 1, we heard more questions like "*is this one?*". In Study 2, we heard more affirmative phrases like "*I think it is this one*". Some participants expressed their positive feeling using *social/emotional* utterances, e.g. "*the head is cute*", "*wow, this is great*", etc. We also heard some chuckles during the study with visual cues. The percentage of *social/emotional* utterance did not significantly increase. This could be because the participants who expressed more feelings were generally more talkative, so they also communicated other things more.

In accordance with their verbal behavior and the previously discussed task completion time, participants in general moved actively and confidently with visual cues. However, for some participants, there were a couple of confusion moments when the bricks were very close but the virtual communication cues were slightly drifting. Overall, the con-

versation in Study 2 was more natural and the participants' movements were more smooth.

8 Design implications and limitations

From these studies we can suggest design implications for future MR remote collaboration systems:

1. Spatial auditory cues give intuitive guidance that can help navigation and object search as well as enhance a user's spatial awareness compared with non-spatial auditory cues.
2. When using spatial auditory cues in a large space, the designer should consider the size of the environment to guarantee the clarity of the rendered sound.
3. Integrating the remote expert's head frustum into the spatial auditory cues can provide significantly better social presence, spatial awareness, and system usability.

However, there are some aspects of our system which could be improved and some limitations with our studies. As we focused the exploration on the local side with a trained remote expert, we did not investigate the remote users' experience with our system. For real-world applications using our system, we may also spatialize the local worker's voice to the remote side and live stream the local environment instead of using a pre-modeled 3D mesh.

We did not include a visual-only condition in Study 2 considering some implementation and evaluation details, but including a visual-only condition might provide more insights into the combined effects of auditory and visual cues. More work could be conducted in this area in future studies.

9 Conclusion and future work

In this paper, we presented a MR remote collaboration system that features both spatial auditory and visual cues. We found that the spatial auditory cues could navigate a local worker in a large space and give the worker better spatial awareness in a MR remote collaborative search task. We also found that the visual cues, especially the head frustum, further helped the local workers to complete the search task faster with a better spatial and social experience. The results of our studies also provide some insights about how the local-remote conversation might become more interactive and intuitive as well as how the local workers could conduct the task more confidently with the integration of spatial auditory and visual cues. Since visual cues are intuitive to perceive and spatial auditory cues indicate 360° directions and distances, the combination of both made some participants feel like they were interacting with a real person.

This work serves as an initial exploration of spatial auditory cues and their combination with visual cues in MR remote collaboration. Section 8 has indicated some directions of future exploration. In addition to these topics, we are also interested in investigating the system performance with more complex tasks (e.g. assembly) or multiple local/remote users.

References

1. Adcock M, Anderson S, Thomas B (2013) RemoteFusion: real time depth camera fusion for remote collaboration on physical tasks. In: Proceedings of the 12th ACM SIGGRAPH international conference on virtual-reality continuum and its applications in industry. pp 235–242
2. Alem L, Li J (2011) A study of gestures in a video-mediated collaborative assembly task. Adv Hum-Comput Interact 2011:1
3. Bai H, Sasikumar P, Yang J, Billinghurst M (2020) A user study on mixed reality remote collaboration with eye gaze and hand gesture sharing. In: Proceedings of the 2020 CHI conference on human factors in computing systems, pp 1–13
4. Barde A, Ward M, Helton WS, Billinghurst M, Lee G (2016) Attention redirection using binaurally spatialised cues delivered over a bone conduction headset. In: Proceedings of the human factors and ergonomics society annual meeting, vol 60. SAGE Publications Sage CA, Los Angeles, pp 1534–1538
5. Beck S, Kunert A, Kulik A, Froehlich B (2013) Immersive group-to-group telepresence. IEEE Trans Visual Comput Graph 19(4):616–625
6. Billinghurst M, Bowskill J, Jessop M, Morphett J (1998) A wearable spatial conferencing space. In: Digest of papers. Second International Symposium on Wearable Computers (Cat. No. 98EX215). IEEE, pp 76–83
7. Billinghurst M, Kato H (2002) Collaborative augmented reality. Commun ACM 45(7):64–70
8. Blessenohl S, Morrison C, Criminisi A, Shotton J (2015) Improving indoor mobility of the visually impaired with depth-based spatial sound. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 26–34
9. Brewster S, Walker V (2000) Non-visual interfaces for wearable computers. In: IEE Colloquium (Digest). IEE; 1999, p 6
10. Brooke J (2013) SUS: a retrospective. J Usability Stud 8(2):29–40
11. Brooke J et al (1996) SUS-a quick and dirty usability scale. Usability Eval Ind 189(194):4–7
12. Buxton W, Moran T (1990) Europarc's integrated interactive intermedia facility (IIIF): early experiences. In: Multi-user interfaces and applications, vol 11, p 34
13. Chen H, Lee AS, Swift M, Tang JC (2015) 3D collaboration method over hololens^{TM} and skype^{TM} end points. In: Proceedings of the 3rd international workshop on immersive media experiences. ACM, pp 27–30
14. DeVincenzi A, Yao L, Ishii H, Raskar R (2011) Kinected conference: augmenting video imaging with calibrated depth and audio. In: Proceedings of the ACM 2011 conference on computer supported cooperative work, pp 621–624
15. Fussell SR, Kraut RE, Siegel J (2000) Coordination of communication: effects of shared visual context on collaborative work. In: Proceedings of the 2000 ACM conference on computer supported cooperative work, pp 21–30
16. Gao L, Bai H, Piumsomboon T, Lee G, Lindeman RW, Billinghurst M (2017) Real-time visual representations for mixed reality remote collaboration

17. Gauglitz S, Nuernberger B, Turk M, Höllerer T (2014) World-stabilized annotations and virtual scene navigation for remote collaboration. In: Proceedings of the 27th annual ACM symposium on user interface software and technology, pp 449–459
18. Gross M, Gross M, Würmlin S, Naef M, Lamboray E, Spagno C, Kunz A, Koller-Meier E, Svoboda T, Van Gool L et al (2003) blue-c: a spatially immersive display and 3D video portal for telepresence. In: ACM Transactions on Graphics, vol 2. ACM, pp 819–827
19. Gupta K, Lee GA, Billinghurst M (2016) Do you see what I see? The effect of gaze tracking on task space remote collaboration. IEEE Trans Visual Comput Graph 22(11):2413–2422
20. Härmä A, Jakka J, Tikander M, Karjalainen M, Lokki T, Hiipakka J, Lorho G (2004) Augmented reality audio for mobile and wearable appliances. J Audio Eng Soc 52(6):618–639
21. Harms C, Biocca F (2004) Internal consistency and reliability of the networked minds measure of social presence
22. Higuch K, Yonetani R, Sato Y (2016) Can eye help you? Effects of visualizing eye fixations on remote collaboration scenarios for physical tasks. In: Proceedings of the 2016 CHI conference on human factors in computing systems. ACM, pp 5180–5190
23. Huang W, Kim S, Billinghurst M, Alem L (2018) Sharing hand gesture and sketch cues in remote collaboration. J Visual Commun Image Represent 58:428–438
24. Joachimczak M Liu J, Ando H (2018) Downsizing: the effect of mixed-reality person representations on stress and presence in telecommunication. In: 2018 IEEE international conference on artificial intelligence and virtual reality, pp 140–143
25. Kim S, Lee G, Sakata N, Billinghurst M (2014) Improving co-presence with augmented visual communication cues for sharing experience through video conference. In: 2014 IEEE international symposium on mixed and augmented reality, pp 83–92
26. Kim S, Lee GA, Sakata N, Dünser A, Vartiainen E, Billinghurst M (2013) Study of augmented gesture communication cues and view sharing in remote collaboration. In: 2013 IEEE international symposium on mixed and augmented reality, pp 261–262
27. Koleva N, Hoppe S, Moniri MM, Staudte M, Bulling A (2015) On the interplay between spontaneous spoken instructions and human visual behaviour in an indoor guidance task. In: CogSci
28. Lee GA, Teo T, Kim S, Billinghurst M (2017) Mixed reality collaboration through sharing a live panorama. In: SIGGRAPH Asia 2017 mobile graphics & interactive applications, pp 1–4
29. Lee GA, Teo T, Kim S, Billinghurst M (2018) A user study on mr remote collaboration using live 360 video. In: 2018 IEEE international symposium on mixed and augmented reality, pp 153–164
30. Loomis JM, Golledge RG, Klatzky RL (1998) Navigation system for the blind: auditory display modes and guidance. Presence 7(2):193–203
31. Müller R, Helmert JR, Pannasch S (2014) Limitations of gaze transfer: without visual context, eye movements do not to help to coordinate joint action, whereas mouse movements do. Acta Psychol 152:19–28
32. Orts-Escolano S, Rhemann C, Fanello S, Chang W, Kowdle A, Degtyarev Y, Kim D, Davidson PL, Khamis S, Dou M et al (2016) Holoportation: virtual 3D teleportation in real-time. In: Proceedings of the 29th annual symposium on user interface software and technology. ACM, pp 741–754
33. Piumsomboon T, Lee GA, Hart JD, Ens B, Lindeman RW, Thomas BH, Billinghurst M (2018) Mini-me: an adaptive avatar for mixed reality remote collaboration. In: Proceedings of the 2018 CHI conference on human factors in computing systems. ACM, New York, NY, USA, pp 46:1–46:13
34. Sandra HG (2006) Development of NASA TLX: result of empirical and theoretical research. San Jose State University, California
35. Sawhney N, Schmandt C (2000) Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. ACM Trans Comput-Hum Interact 7(3):353–383
36. Smith HJ, Neff M (2018) Communication behavior in embodied virtual reality. In: Proceedings of the 2018 CHI conference on human factors in computing systems pp 1–12
37. Sodnik J, Tomazic S, Grasset R, Duenser A, Billinghurst M (2006) Spatial sound localization in an augmented reality environment. In: Proceedings of The 18th Australia conference on computer–human interaction: design: activities, artefacts and environments, pp 111–118
38. Sousa M, dos Anjos RK, Mendes D, Billinghurst M, Jorge J (2019) Warping deixis: distorting gestures to enhance collaboration. In: Proceedings of the 2019 CHI conference on human factors in computing systems. ACM, New York, NY, USA, pp 608:1–608:12
39. Tang TJ, Li WH (2014) An assistive eyewear prototype that interactively converts 3D object locations into spatial audio. In: Proceedings of the 2014 ACM international symposium on wearable computers, pp 119–126
40. Tecchia F, Alem L, Huang W (2012) 3D helping hands: a gesture based MR system for remote collaboration. In: Proceedings of the 11th ACM SIGGRAPH international conference on virtual-reality continuum and its applications in industry, pp 323–328
41. Teo T, Lawrence L, Lee GA, Billinghurst M, Adcock M (2019) Mixed reality remote collaboration combining 360 video and 3D reconstruction. In: Proceedings of the 2019 CHI conference on human factors in computing systems. ACM, New York, NY, USA, pp 201:1–201:14
42. Teo T, Lee GA, Billinghurst M, Adcock M (2018) Hand gestures and visual annotation in live 360 panorama-based mixed reality remote collaboration. In: Proceedings of the 30th Australian conference on computer–human interaction, pp 406–410
43. Tikander M, Karjalainen M, Riikonen V (2008) An augmented reality audio headset. In: Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08), Espoo, Finland
44. Tomczak M, Tomczak E (2014) The need to report effect size estimates revisited. An overview of some recommended measures of effect size. Trends Sport Sci 21(1):19–25
45. Velamkayala ER, Zambrano MV, Li H (2017) Effects of hololens in collaboration: a case in navigation tasks. In: Proceedings of the human factors and ergonomics society annual meeting, vol 61. SAGE Publications Sage CA: Los Angeles, CA, pp 2110–2114
46. Villegas J, Cohen M (2010) "Gabriel": geo-aware broadcasting for in-vehicle entertainment and localizability. In: AES 40th International Conference
47. Vorderer P, Wirth W, Gouveia FR, Biocca F, Saari T, Jäncke L, Böcking S, Schramm H, Gysbers A, Hartmann T et al (2004) Mec spatial presence questionnaire. Retrieved 18 Sept 2015
48. Walker A, Brewster S (2000) Spatial audio in small screen device displays. Pers Technol 4(2–3):144–154
49. Walker A, Brewster S, McGookin D, Ng A (2001) Diary in the sky: a spatial audio display for a mobile calendar. In: People and computers XV—interaction without frontiers. Springer, pp 531–539
50. Yang J, Frank Y, Sörös G (2019) Hearing is believing: synthesizing spatial audio from everyday objects to users. In: Proceedings of the 10th augmented human international conference. ACM, p 28
51. Yang J, Sörös G (2018) Spatial audio for human-object interactions in small AR workspaces. In: Proceedings of the 16th annual international conference on mobile systems, applications, and services. ACM, p 518