Exploring Mixed Reality Robot Communication
Under Different types of Mental Workload

Nhan Tran
Colorado School of Mines
Department of Computer Science
nttran@mines.edu

Kai Mizuno
Colorado School of Mines
Department of Computer Science
kmizuno@mines.edu

Trevor Grant
University of Colorado
Department of Computer Science
trevorgrant@colorado.edu

Thao Phung
Colorado School of Mines
Department of Computer Science
thaophung@mines.edu

Leanne Hirshfield
University of Colorado
Department of Computer Science
leanne.hirshfield@colorado.edu

Thomas Williams
Colorado School of Mines
Department of Computer Science
twilliams@mines.edu

ABSTRACT
This paper explores the tradeoffs between different types of mixed reality robotic communication under different levels of user workload. We present the results of a within-subjects experiment in which we systematically and jointly vary robot communication style alongside level and type of cognitive load, and measure subsequent impacts on accuracy, reaction time, and perceived workload and effectiveness. Our preliminary results suggest that although humans may not notice differences, the manner of load a user is under and the type of communication style used by a robot they interact with do in fact interact to determine their task effectiveness.

KEYWORDS
augmented reality, mixed reality, cognitive load, deictic gesture, human-robot interaction

Due to the near-necessity of deictic gestures in situated communication, human-robot interaction researchers have sought to enable robots to understand [28] and generate [38–40] deictic gestures just as humans do. But while understanding deictic gestures requires only a camera or depth sensor, generation of deictic gestures requires a specific robotic morphology (i.e., expressive robotic arms). This fundamentally limits the gestural capabilities, and thus overall communicative capabilities, of the majority of robotic platforms in use today, such as mobile bases used in warehouses, assistive wheelchairs, and unmanned aerial drones. Moreover, even for robots that do have arms, traditional deictic gestures have fundamental limitations. In contexts such as urban or alpine search and rescue, for example, robots may need to communicate about hard-to-describe and/or highly ambiguous referents in novel, uncertain, and unknown environments.

To demonstrate all of these problems, consider an aerial drone in a search and rescue context that needs to generate an utterance such as "I found a victim behind that tree" [cf.64]. First, the robot is highly unlikely to have an arm mounted to it, and thus physical gesture is simply not a possibility. Second, even if the robot did somehow have an arm mounted on it, a pointing gesture is unlikely to be able to successfully pick out a specific far-off tree, and the natural language needed to disambiguate it is likely to be either extremely complex ("the fourth tree from the left in the clump of trees to the right of the large boulder") or non-human-understandable ("the tree 48.2 meters in that direction").

To address these limitations of traditional "egocentric" physical gestures, researchers have recently been exploring the use of mixed reality deictic gestures [63]: visualizations that can serve the same purpose as traditional deictic gestures, and which fall within the broad category of view-augmenting mixed reality interaction design elements in the Reality-Virtuality Interaction Cube framework of Williams, Szafir, and Chakraborti [60]. Williams et al. [63] divides these new forms of non-egocentric visual gestures into allocentric visualizations that can be displayed in teammates’ augmented reality head-mounted displays, and perspective-free visualizations that can be projected onto the ground. Recent work in this space has focused on allocentric gestures such as circles and arrows drawn over target objects [57, 58], as well as ego-sensitive allocentric gestures such as virtual arms [17, 35]. Williams et al. [58], for example [see also 57], demonstrate that (non-ego-sensitive) allocentric virtual gestures, at least when tested in a simulated video-based experiment, have

1 INTRODUCTION
This paper explores the tradeoffs between different types of mixed reality robotic communication under different levels of user workload.

Successful human-robot interaction in many domains relies on successful communication. Accordingly, there has been a wealth of research on enabling human-robot communication through natural language [29, 51]. However, just like human natural language communication, situated human-robot dialogue is inherently multimodal, and necessarily involves communication channels other than speech. For a host of reasons both egocentric (sensitive only to their own perspective) and allocentric (sensitive to others’ perspectives), people regularly use gaze and gesture cues to augment, modify, or replace their natural language utterances. Speakers regularly use deictic gestures such as pointing, for example, to direct interlocutors’ attention to objects in the environment, both to reduce the number of words that the speaker must use to refer to their target referents as well as to lower the cognitive burden of listeners in interpreting speakers’ utterances.
we conclude with a discussion of our results and directions for future work. Moreover, as previously pointed out by Hirshfield et al. [21], the tradeoffs previously considered by Williams et al. [58] between language and visual gesture may be highly sensitive to the level and type of cognitive load that teammates are under. For example, Hirshfield et al. [21] suggest that it may not be advantageous to rely heavily on visual communication in contexts with high visual load (or to rely heavily on linguistic communication in contexts with high auditory or working memory load). These intuitions are motivated by prior theoretical work on human information processing, including Multiple Resource Theory [56], the Perceptual Load model [26], and the Dual-Target Search model [33].

In this paper we thus also present the first exploration of the tradeoffs between different forms of mixed reality communication in contexts with different types of workload impositions. The rest of the paper will proceed as follows. In Section 2, we discuss some additional important related work related to both AR-for-HRI and cognitive load estimation. In Section 3, we present the design of a human-subject experiment that uses this technical approach to study the effectiveness of different robot communication styles under different types of cognitive load. In Section 4, we present the results of this experiment. Finally, in Sections 5 and 6 we conclude with a discussion of our results and directions for future work.

2 RELATED WORK

2.1 AR for HRI

Mixed reality technologies that integrate virtual objects into the physical world have recently sparked research interest in the Human-Robot Interaction (HRI) community [61] because they enable better exchange of information between people and robots, in order to improve shared mental models, calibrated trust, and situation awareness [49].

While there has been significant research on augmented and mixed reality for several decades, [3, 4, 7, 52, 65] and acknowledgement of the potential for impact of AR on HRI for many years as well [16, 31], it is only in recent years that there has been significant and sustained interest in AR-for-HRI [61, 62]. Recent works in this area include approaches using AR for robot design [34], calibration [42], and training [46]. Moreover, there are a number of approaches towards communicating robots’ perspectives [19], intentions [2, 9, 10, 12, 15], and trajectories [14, 30, 36, 54].

One of the best ways to improve human robot interaction is sharing perspectives of human and robot to each other. Amor et al. [1] suggest that projecting human instructions and robot intentions (by highlighting potential target objects) in a constrained and highly structured task environment improves human robot teamwork and produces better task result [1, 2, 15]. Similarly, Sibirtseva et al. [44], present work in which robots receiving natural language instructions reflexively generate augmented reality annotations surrounding candidate referents as they are being disambiguated [44].

Finally, as described above, in our own work, we have investigated the use of AR augmentations as an active rather than passive communication strategy, generated as gestures accompanying natural language communication [35, 57, 58].

2.2 Objective Measurements of Cognitive Load

Hirshfield et al. [21] suggest several contextual factors that may determine when mixed reality deictic gestures are most helpful to human teammates: teammates’ cognitive load may dictate whether they are capable of accepting new information; and their auditory and visual perceptual load may dictate the most effective modality to accompany or replace natural language. These neural correlates of cognitive and perceptual states can be collected in real-time using neurological and physiological sensors for unobtrusively measuring humans’ brain and physiological data including functional Near-Infrared Spectroscopy (fNIRS), Electroencephalography (EEG), electrodermal activity (EDA), electrocardiogram (ECG), and Respiration sensors [13]. fNIRS, a lightweight and non-invasive device, is gaining popularity in the Human-Computer Interaction community [45], as it offers several advantages over other braincomputer interface (BCI) technologies such as greater spatial resolution, higher signal-to-noise ratio, and better practicality for use in normal working conditions [20, 43], although it is of course subject to other limitations [47, 48], including in HRI contexts [8].

The fNIRS component handles raw data from the sensor and outputs a multilabel vector consisting of three labels (workload, auditory perceptual load, and visual perceptual load) from a multilabel long short-term memory (LSTM) classifier every second. These labels are sent to and processed by the centralized server, which then communicates the appropriate decision to both the robot and the mixed reality headset. In this work we are not yet using the fNIRS component of our architecture, and are instead using experimental manipulation to systematically vary cognitive load.

3 EXPERIMENT

In this section we present the design of a human-subject experiment that uses this technical approach to study the effectiveness of different robot communication styles under different types of cognitive load.

3.1 Hypotheses

Specifically, this experiment was designed to test the following hypotheses, which formalize the intuitions of Hirshfield et al. [21].

H1 Users under high visual perceptual load will perform best when robots rely on complex natural language without the use of mixed reality deictic gestures.

H2 Users under high auditory perceptual load will perform best when robots rely on mixed reality deictic gestures without the use of complex natural language.

H3 Users under high working memory load will perform best when robots rely on mixed reality deictic gestures without the use of complex natural language.
H4 Users under low overall load will perform best when robots rely on mixed reality deictic gestures paired with complex natural language.

3.2 Task Design

![Figure 1: Our experimental setup.](image)

The user’s primary task is to look out for a particular block in the block panel (selected from among red cube, red sphere, red cylinder, yellow cube, yellow sphere, yellow cylinder, green cube, green sphere, green cylinder1). These nine blocks were formed by combining three colors red, yellow, green with three shapes cube, sphere, cylinder. Whenever they see this target block, their task is to pick-and-place it into any one of a particular set of bins. For example, a user might be told that whenever they see a red cube they should place it in bins two or three.

Two additional factors increase the complexity of this primary task. First, in order to force participants to remember the full set of candidate bins, rather than just one particular bin from that set, at every point during the task one random bin is marked as unavailable (with the disabled bin changing each time a block is placed in a bin). Second, to allow us to examine auditory load, the user hears a series of syllables playing in the task background (selected from among bah, beh, boh, tah, teh, toh, kah, keh, koh). These nine syllables were formed by combining three consonant sounds b,t,k with three vowel sounds ah,eh,oh. The user is given a target syllable to look out for, and told that whenever they hear this syllable, the bins that they should consider to place blocks in should be exchanged with those they were previously told to avoid. For example, if the user’s target bins from among four bins are bins two and three, and they hear the target syllable, then future blocks will need to be placed instead into bins one and four.

Secondary Task

![Figure 2: Experiment in progress.](image)

Three times per experiment block, the participant encounters a secondary task, in which the Pepper robot interjects and asks the participant to move a particular, currently visible block, to a particular, currently accessible bin.

3.3 Experimental Design

We used a Latin square counterbalanced within-subjects experimental design with two independent variables serving as within-subject factors:

![Figure 2: Experiment in progress.](image)

**Cognitive Load**

Our first independent variable, cognitive load was manipulated through our primary task. Following Beck and Lavie [5], we manipulated communication style by jointly manipulating memory constraints and target/distractor discriminability (ep. [25]), producing four different load profiles: one in which all load was considered low; one in which only working memory load was considered to be high, and one in which only visual perceptual load was considered to be high, and one in which only auditory perceptual load was considered to be high.

Working memory load was manipulated as follows: In the high working memory load condition, participants were required to remember the identities of three target bins out of a total of six visible bins, producing a total memory load of seven items when including the two properties of the target block (shape and color) and the two properties of the target syllable (consonant and vowel). In all other conditions, participants were only required to remember the identities of two target bins out of a total of four visible bins, producing a total memory load of six items.

Visual perceptual load was manipulated as follows: In the high visual perceptual load condition, the target block was always difficult to discriminate from distractors due to sharing of one...
common property with all distractors. For example, if the target block was a red cube, all distractors would be either red or cubes (but not both). In the low visual perceptual load condition, the target block was always easy to discriminate from distractors due to sharing no common properties with any distractors. For example, if the target block was a red cube, no distractors would be red or cubes.

Auditory perceptual load was manipulated as follows: In the high auditory perceptual load condition, the target syllable was always difficult to discriminate from distractors due to sharing of one common property with all distractors. For example, if the target syllable was *kah*, all distractors would either start with *k* or end with *ah* (but not both). In the low auditory perceptual load condition, the target syllable was always easy to discriminate from distractors due to sharing no common properties with any distractors. For example, if the target syllable was *kah*, no distractors would either start with *k* or end with *ah*.

**Communication Style**

Our second independent variable, communication style, was manipulated through our secondary task. Following Williams et al. [57] and Williams et al. [58], we manipulated communication style by having the robot exhibit one of three behaviors:

During experiment blocks associated with the complex language communication style condition, the robot with which participants interacted referred to objects using full referring expressions needed to disambiguate those objects.

During experiment blocks associated with the complex language + AR communication style condition, the robot with which participants interacted referred to objects using full referring expressions needed to disambiguate those objects (e.g., “the red sphere”), paired with a mixed reality deictic gesture (an arrow drawn over the object to which the robot was referring).

During experiment blocks associated with the simple language + AR communication style condition, the robot with which participants interacted referred to objects using minimal referring expressions (e.g., “that block”), paired with a mixed reality deictic gesture (an arrow drawn over the object to which the robot was referring).

Following Williams et al. [57] and Williams et al. [58], we did not examine the use of simple language without AR, as that communication style does not always allow complete referent disambiguation, resulting in the user needing to ask for clarification or guess at random between ambiguous options.

**3.4 Measures**

We expected performance improvements to manifest in our experiment in four different ways: task accuracy, task reaction time, perceived mental workload, and perceived communicative effectiveness.

These aspects of performance were measured as follows:

**Accuracy** was measured for both primary and secondary tasks by logging which virtual object participants clicked on, and determining whether or not this was the object intended by the task or by robot.

**Reaction time** was measured for both primary and secondary tasks by logging time stamps at the moment participants interacted with virtual objects (both blocks and bins). In the primary task, reaction time was measured as the time between placement of the previous primary target block and picking of the next primary target block. In the secondary task, reaction time was measured as the time between the start of Pepper’s utterance and the placement of the secondary target block.

**Perceived mental workload** was measured using a NASA Task Load Index (NASA TLX) survey [18] administered at the end of each experiment block.

**Perceived communicative effectiveness** was measured using the modified version of the Gesture Perception Scale [40] previously employed by Williams et al. [57, 58], which was delivered along with the NASA TLX Survey at the end of each experiment block.

**3.5 Procedure**

Upon arriving at the lab, providing informed consent, and completing demographic and visual capability survey, participants were introduced to the task through both verbal instruction and an interactive tutorial.

![Tutorial](image3)

The tutorial scene provides text and visuals that walk the participant through how a round in the experiment will function. When the participant starts the tutorial, they see a panel with text-instructions, a row of blocks, and four bins (Fig. 3). Participants are walked through how to use the HoloLens air tap gesture to pick up blocks and put them in bins through descriptive text and an animation showing an example air tap gesture, and informed of task mechanics with respect to both target/non-target bins and temporarily disabled grey bins. Participants then start to hear syllables being played by the HoloLens. When the target syllable *teh* plays, the target and non-target bins switch. Each bin on screen is labeled as a ‘target’ or ‘non-target’, in order to help the participant understand what is happening when the target syllable plays. These labels are only shown in the tutorial and participants are reminded that they will have to memorize which bins are targets for the actual game. At the end of the tutorial the participant has...
to successfully put a target block in a target bin three times before they can start the experiment.

After completing this experiment, participants engaged in each of the twelve (Latin square counterbalanced) experiment blocks formed by combining the four cognitive load conditions and the three communication style conditions, with surveys administered after each experiment block.

3.6 Participants

36 participants were recruited from Colorado School of Mines (31 M, 5 F), ranging in age from 18 to 32. None had participated in any previous studies from our laboratory.

3.7 Analysis

Data analysis was performed within a Bayesian analysis framework using the JASP 0.11.1 [50] software package, using the default settings as justified by Wagenmakers et al. [53]. For each measure, a repeated measures analysis of variance (RM-ANOVA) [11, 32, 37] was performed, using communication style and cognitive load as random factors. Baws factors [27] were then computed for each candidate main effect and interaction, indicating (in the form of a Bayes Factor) for that effect the evidence weight of all candidate models including that effect compared to the evidence weight of all candidate models not including that effect. When sufficient evidence was found in favor of a main effect, the results were further analyzed using a post-hoc Bayesian t-test [23, 55] with a default Cauchy prior (center=0, r=√2=0.707). When sufficient evidence was found in favor of an interaction effect, the results were further analyzed using a series of post-hoc paired-samples t-tests each category of cognitive load.

4 RESULTS

4.1 Reaction Time

Secondary Task

Figure 4: Effect of communication strategy (complex language + AR vs. complex language vs. simple language + AR) on secondary task reaction time.

Figure 5: Effect of workload (Low All) vs. (High Visual) vs. (High Auditory) vs. (High Working Memory) on participant’s secondary task reaction time.

Figure 6: Effect of both workload and communication strategy on participant’s secondary task reaction time.

Our results provided extreme evidence in favor of effects of both communication style (Bf 3.109e29) and cognitive load (Bf 9.881e9) on secondary task reaction time, as shown in Figs. 4 and 5, as well as an interaction between communication style and cognitive load (Bf 1.160e12) on reaction time, as shown in Fig. 6.

Post-hoc analysis of the main effect of communication style on secondary task reaction time revealed significant differences specifically between the use of complex language alone (μ = 8.116sec) and both complex language + AR (μ = 7.399sec, Bf 2.955e21) and simple language + AR (μ = 7.501sec, Bf 9.396e15), with anecdotal evidence against a difference between complex language + AR and complex language alone (Bf = 0.46 in favor of an effect; 1/0.46 = Bf 2.14 against an effect)

This yields a preference ordering complex language < (simple language + AR = complex language + AR) when cognitive

Bayes Factors above 100 indicate extreme evidence in favor of a hypothesis [6, 22]. Here, for example, our Baws Factor Bf of 7.024e25 suggests that our data were 7.024e25 times more likely to be generated under models in which communication style is included than under those in which it is not.
load is not considered.

Post-hoc analysis of the main effect of cognitive load on secondary task reaction time revealed significant differences specifically between conditions with high auditory perceptual load ($\mu = 7.37$ sec) and all other conditions, i.e., low overall load ($\mu = 7.66$ sec, Bf 2931.437), high visual perceptual load ($\mu = 7.76$ sec, Bf 283407.874), and high working memory load ($\mu = 7.88$ sec, Bf 1.343e9), as well as between conditions with high working memory load and those with low overall load (Bf 13.381).

This yields a preference ordering high auditory perceptual load < (low overall load < high working memory load) = high visual perceptual load when communication style is not considered.

Post-hoc analysis of the interaction effect between communication style and cognitive load on secondary task reaction revealed the following additional findings:

**Low Overall Load** Extreme evidence was found under low overall load between each pair of communication strategies (simple language + AR ($\mu = 7.56$, $SD = 0.732$) vs complex language alone ($\mu = 8.19$, $SD = 0.685$) Bf 8.995e6; simple language + AR vs complex language + AR ($\mu = 7.25$, $SD = 0.654$) Bf 705110.101; complex language alone vs complex language + AR Bf 1.281e13.

This yields a preference ordering complex language alone < simple language + AR < complex language + AR in the low overall load condition.

**High Working Memory Load** Extreme evidence was found under high working memory load between simple language + AR ($\mu = 7.43$, $SD = 0.565$) and both complex language alone ($\mu = 8.24$, $SD = 0.327$ Bf 1.080e7) and complex language + AR ($\mu = 7.98$, $SD = 0.746$ Bf 2076.594).

This yields a preference ordering (complex language alone = complex language + AR) < simple language + AR in the high working memory load condition.

**High Visual Perceptual Load** Moderate to extreme evidence was found under high visual perceptual load between complex language + AR ($\mu = 7.50$, $SD = 0.456$) and both complex language alone ($\mu = 7.99$, $SD = 0.747$ Bf 1449.784) and simple language + AR ($\mu = 7.78$, $SD = 0.508$ Bf 5.336).

This yields a preference ordering (simple language + AR = complex language alone) < complex language + AR in the high visual perceptual load condition.

**High Auditory Perceptual Load** Extreme evidence was found under high auditory perceptual load between each pair of communication strategies (simple language + AR ($\mu = 7.21$, $SD = 0.367$) vs complex language alone ($\mu = 8.05$, $SD = 0.421$) Bf 7.374e6; simple language + AR vs complex language + AR ($\mu = 6.85$, $SD = 0.560$) Bf 35.760; complex language alone vs complex language + AR Bf 1.126e13.

This yields a preference ordering complex language alone < simple language + AR < complex language + AR in the high auditory perceptual load condition.

**Primary Task**

Strong evidence was found against any effects of communication style or cognitive load on primary task reaction time (All Bfs > 20 against an effect).

**4.2 Accuracy**

Strong evidence was found against any effects of communication style or cognitive load on primary or secondary task accuracy (All Bfs > 27 against an effect).

**4.3 Perceived Mental Workload**

Anecdotal to strong evidence was found against any effects of communication style or cognitive load on perceived mental workload (Bfs between 22.43 and 40.91 against an effect).

**4.4 Perceived Communicative Effectiveness**

Anecdotal to strong evidence was found against any effects of communication style or cognitive load on perceived communicative effectiveness (Bfs between 2.23 and 83.33 against an effect on all questions).

**5 DISCUSSION**

Our results suggest that although humans may not be aware of differences in their performance or mental workload when different mixed reality robotic communication styles are used, or when they are under different types of cognitive load, both of these factors do in fact influence the speed at which they are able to accomplish tasks.

First, our results suggest that different types of mental workload do, unsurprisingly, impact task performance, with participants under low overall load reacting more quickly than participants under high working memory load. What is surprising is that participants under high auditory load clearly demonstrated the fastest reaction times overall. It is not yet clear how to interpret this result, but it is possible that this effect is due to individuals generally responding faster to auditory stimuli than visual [24].

Second, our results suggest, unsurprisingly, that different communication strategies impact task performance. In fact, our results exactly match what we observed in previous experiments [58]: participants demonstrate slower reaction times when complex language alone is used, with no clear differences between simple and complex language when it is augmented with a mixed reality deixtic gesture.

Finally, our results suggest a complex interplay between communication style and cognitive load. Specifically, our results suggest that while using complex language + AR resulted in the best task performance in most workload conditions (an encouraging result given that our previous work has shown that participants find robots most likeable when they use this communication style), this does not hold true when users are under high working memory load. Rather, when users are under high working memory load, it is best to use simple language + AR, to avoid overloading participants.
Overall, these results support hypotheses H3 and H4, but fail to support hypotheses H1 and H2. While our original expectation was that the differences between communication styles under different cognitive load profiles would primarily be grounded in whether communication style was overall visual or overall visual, in fact what we observed is that visual augmentations are always helpful, and differences in effectiveness between workload conditions depend entirely on whether or not the user is under high cognitive load.

While we observed clear impacts of workload profiles on task performance, participants did not demonstrate any differences in perceived workload or perceived effectiveness. This could be the case that the differences in reaction time simply were not large enough for participants to notice: the observed differences were on the order of one second of reaction time when overall reaction time was around 7.5 seconds. Participants may simply not have noticed a 15% speed increase in certain conditions, or may not have attributed it to the robot.

This could also be the case due to overall task difficulty. While participants’ TLX scores had a mean value of approximately 21 out of 42 points in all conditions (i.e., the data was nearly perfectly centered around “medium” load), analysis of individual performance trajectories demonstrates that the task was sufficiently difficult that many participants experienced catastrophic primary task shedding, often immediately after a primary task (likely due to missing an auditory cue while dealing with a secondary task). As illustrated in Fig. 7, task performance varied significantly between participants. In this figure the dark X markers represent the time the robot started uttering a secondary task requests. As can be seen, most participants performed well on the primary task (resulting in many green dots) up until immediately after the first or second secondary task request. As can also be seen, when participants made a mistake, except in cases where the error fell between secondary task initiation and completion, they often failed to recover from the failure. In future work, we hope to conduct deeper analyses of these trajectories, specifically examining factors such as differences in task completion speeds and rates of catastrophic primary task shedding across conditions.

6 CONCLUSION

The ultimate goal of our research is to enable adaptive mixed reality communication for human-robot interaction. In this paper, we presented the first experimental steps towards achieving this goal. Our results provide critical insights for the future design of our proposed adaptive system.

In future work, we plan to complete our integration of the fNIRS neurophysiological sensor with the current mixed reality robotic architecture, in order to accurately measure changes in mental workload within experimental conditions, as well as in task contexts that do not have tightly controlled levels of workload. We further plan to integrate all three components together with the Distributed Integrated Affect Reflection and Cognition (DIARC) architecture to

---

3This figure shows only one condition, the complex language/High visual perceptual load condition, for the sake of space. All twelve condition plots, however, show similar results to what we observed here.
leverage its rich natural language understanding and generation capabilities [41, 59].

Finally, in future work we also hope to consider how robots can tailor gestural cues to be easily discriminable from both background visual stimuli and other task targets without placing the human teammate at risk of inattentional blindness.

ACKNOWLEDGEMENT

This research was funded in part by NSF grant IIS-1909864.

REFERENCES


[50] JASP Team. 2018. JASP (Version 0.8.5.1)[Computer software].


