Augmented Reality Dialog Interface for Multimodal Teleoperation

André Pereira, Elizabeth J. Carter, Iolanda Leite, John Mars, Jill Fain Lehman[1]

***Abstract*— We designed an augmented reality interface for dialog that enables the control of multimodal behaviors in telepresence robot applications. This interface, when paired with a telepresence robot, enables a single operator to accurately control and coordinate the robot's verbal and nonverbal behaviors. Depending on the complexity of the desired interaction, however, some applications might benefit from having multiple operators control different interaction modalities. As such, our interface can be used by either a single operator or pair of operators. In the paired-operator system, one operator controls verbal behaviors while the other controls nonverbal behaviors. A within-subjects user study was conducted to assess the usefulness and validity of our interface in both single and paired-operator setups. When faced with hard tasks, coordination between verbal and nonverbal behavior improves in the single-operator condition. Despite single operators being slower to produce verbal responses, verbal error rates were unaffected by our conditions. Finally, significantly improved presence measures such as mental immersion, sensory engagement, ability to view and understand the dialog partner, and degree of emotion occur for single operators that control both the verbal and nonverbal behaviors of the robot.**

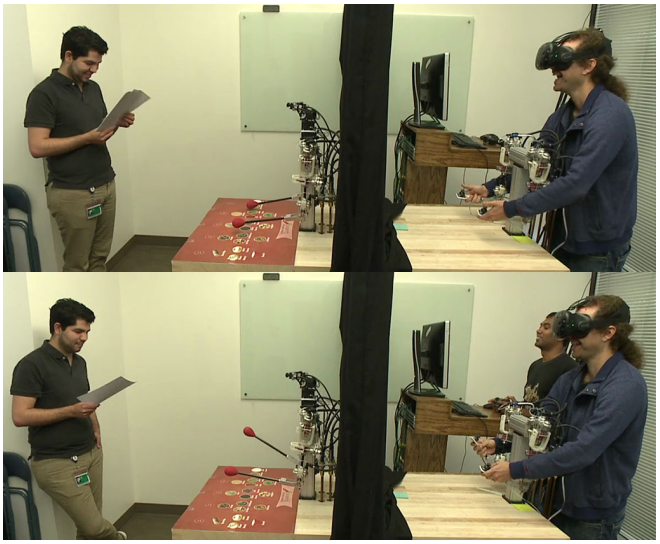

Fig. 1: A single operator (top) and two operators (bottom).

I. Introduction

Recent advances in Virtual Reality (VR) and Augmented Reality (AR) technology have significantly improved user experience in a variety of devices and applications. In this work, we explore the use of this technology to improve telepresence robotic interfaces. More specifically, we designed and implemented an AR interface that, when paired with a telepresence platform, enables a single operator to control the language, gaze, and body movements of a robot without the need for pre-programmed sets of nonverbal behaviors. Such interfaces are important for the domain of robot telepresence, to control robots in Wizard of Oz (WoZ) experiments or entertainment venues (e.g., theme parks), and to collect training data.

In a prior user study, the same telepresence platform required two operators to control the robot's behavior [1]. One operator wore a VR headset and controlled the robot's nonverbal behavior (gaze and gestures) while the other was responsible for the verbal behavior of the robot (text-to-speech output). On the one hand, this kind of setup causes coordination issues because the operator responsible for the robot's nonverbal behavior is not aware of the exact timing of the verbal output controlled by the second operator. On the other hand, a single operator controlling all modalities might experience higher cognitive load, which can result in slower response times or mistakes. To assess the advantages of both setups, our AR interface allows the teleoperation of a robot by a single operator or by a pair of operators (Figure 1). Solo operators control the verbal behaviors by wearing a VR headset that overlays the video feed from the robot's perspective with our AR interface. Joysticks attached to the arms of our telepresence robot enable interface control and leave users' hands in the optimal locations to perform gestures by moving the robot's arms. By freely moving their heads, users can also control the robot's gaze behaviors. When two operators are involved in the interaction, one wears the VR headset and is solely responsible for controlling the nonverbal behavior. A second operator looks at a monitor that displays the same AR interface and manipulates it by using a game controller. The shared AR interface allows for a beneficial feedback loop between operators.

We conducted a within-subjects study in which we invited pairs of people to operate the robot individually and together. The role of the participants was to control a believable and responsive robotic fast-food waiter that could interpret and react to complex and unexpected situations delivered by an actor. We recorded the performance of the robot as it conversed with the actor and analyzed the data to assess the number of errors, response times, and coordination offsets. Presence questionnaires were administered after each condition and at the end of the experiment. Our proposed AR dialog interface and the results obtained in this study have relevant implications for the future design of interfaces for teleoperated robot systems.

[1]All authors are with Disney Research, Pittsburgh, USA jill.lehman@disneyresearch.com

II. Related Work

Commercially available VR devices have increased the appearance of virtual and augmented reality applications in several domains [2], [3]. In Human-Robot Interaction (HRI), the most common use case for the technology has been remote, first-person robot teleoperation. One of the first instances was proposed by Martins and Ventura [4], who implemented a head-tracking system using a head-mounted display for controlling a search-and-rescue mobile robot. Similar to our telepresence platform, their system enables a person to control the robot's stereo vision module such that the cameras mounted in the robot follow the controller's head movements. Several authors have explored the use of head-mounted displays for controlling navigation versus navigation by game controllers, particularly for Unmanned Aerial Vehicles (UAV) [5]. For example, Pittman and LaViola [6] compared multiple head-tracking techniques with traditional game controllers in a UAV navigation task. Despite an overall preference for the game controllers, participants considered head rotation a fun and responsive technique for controlling the robot.

To improve the user's sense of immersion, Kratz et al. [7] investigated the effect of stereo versus non-stereo vision and low versus high camera placement for improving the video feed of a head-mounted display in a mobile robot telepresence scenario. The results of a pilot study suggest that users preferred the non-stereo vision, regardless of the camera placement. More recently, Fritsche et al. [8] presented the first teleoperation system to enable first-person control of a humanoid robot using a head-mounted display, skeleton tracking, and a glove that provides haptic feedback to the operator. They report that a human operator could use their system to successfully complete imitation and pick-and-place tasks via an iCub robot.

With the exception of the work by Fritsche [8], the other VR systems presented here allowed the control of a single modality. In HRI applications, however, experimenters often need to control multiple robots [9] or different modalities in parallel. A recent survey showed that more than half of the papers published in the HRI conference reported studies where a human controlled at least one of the robot's functions [10]. In an earlier review, Riek [11] found that the most common tasks performed by a human controlling a robot in WoZ settings were natural language processing and nonverbal behavior. Despite the prominence of WoZ studies in HRI, most tools are created *ad hoc* for a particular experiment or robot. However, a few authors have proposed generic, configurable WoZ interfaces. Some examples are *Polonius* [12], a modular interface for the ROS-Framework; *DOMER* [13], a tool for controlling a Nao robot in therapeutic settings; and *OpenWoZ* [14], a web-based architecture.

The contributions of the current work are twofold. First, we extend prior research by proposing an AR interface that enables a single operator to control the language, gaze and body movements of a robot. Second, we compare measures of task performance and presence for one vs. two operators.

III. Hardware Setup

We use Jimmy, a hybrid hydrostatic transmission and human-safe haptic telepresence robotic platform [15]. The platform contains two upper body torsos with four degrees of freedom (DOF) in each arm. The four DOF allow operators both to feel haptic feedback and to freely move the operator's side of the platform's arms while the movement is fully replicated on the opposite side. The viewer's side of the platform contains two cameras mounted on a 2-DOF neck controlled by robotic servos. The cameras are used to stream real-time video to an operator's headset that also maps the head orientation of the operator to the neck servos of the robot. A pair of speakers is used for the output of Jimmy's text-to-speech. Using this platform and a curtain between the human operator and the viewer, the operator can experience the interaction and environment from the robot's perspective while the viewer experiences believable, noiseless, and fluid nonverbal behaviors on the other side. Figure 1 shows both the Solo and Pair setups. The Solo setup is composed of Jimmy, a VR headset, and two analog 2-axis thumb joysticks located at the extremities of Jimmy's "arms". Using this system, the operator can control the verbal overlay interface described in the next subsection while simultaneously directing the robot's nonverbal behaviors. In the Pair setup, two human operators control different aspects of the robot; one controls gaze and gesture movements (Nonverbal Operator), while the second controls Jimmy's speech (Verbal Operator). In the Pair configuration, the joysticks on Jimmy's arms are deactivated and their function replaced with a game controller with two analog joysticks of similar size. The Verbal Operator looks at a computer monitor that mirrors the image displayed in the headset and uses the controller to manipulate the dialog interface.

Additionally, two high-definition video cameras were used to record the interactions. One of the cameras was placed from the actor's perspective (i.e., viewing the robot and restaurant menu face-to-face as shown in Figure 3) and another captures a side view (i.e., viewing the operator(s) as well as the robot and actor, as shown in Figure 1). The restaurant menu for the actor and the robot to share was placed on the table immediately in front of the robot.

IV. Augmented Reality Interface

To engage in task dialog, we designed an interface that overlays the real image captured by the robot's camera(s) with contextually-determined dialog choices (Figure 2). This creates an augmented reality environment that allows one or two operators to control a robot's verbal and nonverbal behavior. The camera input is integrated into the Unity3D game engine, and we use its GUI design capabilities to draw the interface. Steam's OpenVR SDK then transfers camera frames into the headset. In the Pair conditions, a full-screen copy of the interface is displayed on a desktop monitor. To facilitate conversational grounding [16] and create a fair comparison of operators' performances, we keep the user interface constant across Solo and Pair conditions.

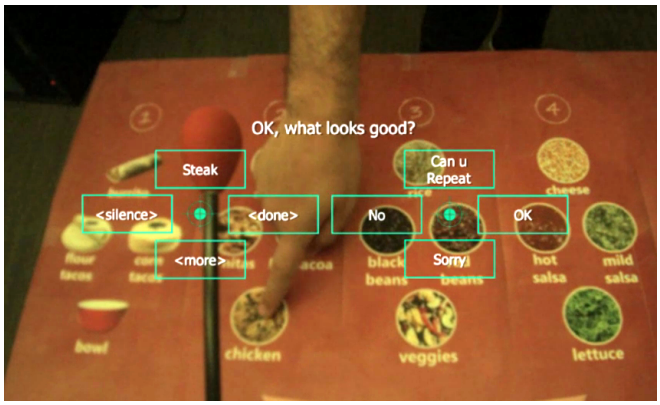

Fig. 2: The verbal overlay interface.

The interface has two distinct areas with multiple options each. The Right Hand Menu (RHM) gives the operator a limited number of options for direct language initiative, mainly in the service of conversational repair. For our task, the RHM does not change and contains "No", "OK", "Say that again", and "Sorry" (Figure 2, right). The Left Hand Menu (LHM) allows the operator to perform the robot's language-understanding function by mapping the customer's behavior into the small set of context changes or utterance categories expected by the operator at each point in the ordering process (Figure 2, left). When the operator selects an LHM option that describes the user's current behavior, the contents of the LHM change automatically according to the predefined dialog model. Note that one position in the LHM in the figure is labeled "more," giving the operator access to additional possible mappings in a sub-menu. Because the LHM presents the operator with options that are relevant in the current moment, external communication from different sensors and inputs could be used to automate the choices in this part of the interface further.

In physical terms, the operator selects from either menu by using the corresponding joystick (left or right) to touch a crosshair icon to the option for a fixed period. Initially, operators had to move the joystick and click to select, but participants in pilot testing found that hovering for a fixed period seemed to reduce errors and be more intuitive. When the crosshair touches an option, it highlights that option and gives visual feedback to the operator(s) that the choice is about to be selected. This kind of feedback is of particular importance in providing opportunities for nonverbal coordination when two operators are controlling the robot; the nonverbal operator can use it to predict which option is going to be picked up by the verbal operator. Single operators also can use this information to coordinate the robot's nonverbal behaviors while the robot uses its text-to-speech engine to communicate verbally. Note additionally that when the operator makes a selection in the LHM that is tied to an output utterance in the dialog model, a subtitle label immediately appears in the display to give the operator information about what the robot will say.

Given this interface and control method, we could, in theory, present up to eight options for each hand in 45-degree increments around the center point of the joystick, with four options in the vertical and horizontal extremities and four in the diagonal positions. However, we decided to use only four options in each area for this experiment to force the use of sub-menus, which could lead to errors and coordination issues during the interaction. Although not strictly necessary for our task, multi-level menus would typically be necessary for scenarios with more complex dialog and longer time periods available for training.

V. Evaluation

In the user study, we measured participants' performance and subjective personal experiences when controlling each of the two versions of the robot teleoperation we described earlier.

A. Participants

We recruited 16 pairs of adult participants: 9 female, 23 male, ages 21-37 years, with mean(stdev) age = 28.8(4.2). We verified that all pairs consisted of two friends so that the level of comfort with the other operator in a pair would not present confounds. Participants were recruited via email lists and word of mouth. We verified that they had normal or corrected-to-normal hearing and vision and no mobility problems that would affect their performance in the experiment. Our Institutional Review Board approved this research and participants were paid for their time.

B. Scenario

Interactions consisted of scenarios in which a customer (played by an actor) enters a build-your-meal restaurant, *Jimmy's Mexican Eatery*, and the teleoperated robot serves him. The customer had a restaurant menu with various types of items from which to choose at each phase (i.e., meal type, protein options, carbohydrate options, toppings, and special additions). The robot operator(s) must greet the customer, obtain the order, tap each selected item, and bid farewell to the customer at the end of the interaction. Figure 3 shows the view from the actor's perspective.

During the ordering process, the robot repeats back each individual item selected. Repetition encourages accuracy by providing feedback that makes any error in the selection process more perceptible. Additionally, to present coordination challenges, participants were also instructed that the robot should tap the items on Jimmy's restaurant menu to record the order. These two reactions added verisimilitude to the interaction, as they often happen in restaurants to ensure that the server has heard the customer correctly and is choosing the appropriate items.

1) Scripts: Three sets of testing scripts defined the interactions between the actor and the robot. Each set contained both an Easy and a Hard script. Easy scripts contained 11 instances of turn-taking (e.g., see Table I), matched for the customer language phenomena (ordering items at Level 1 of the LHM, ordering Level 2 items, a turn that necessitates

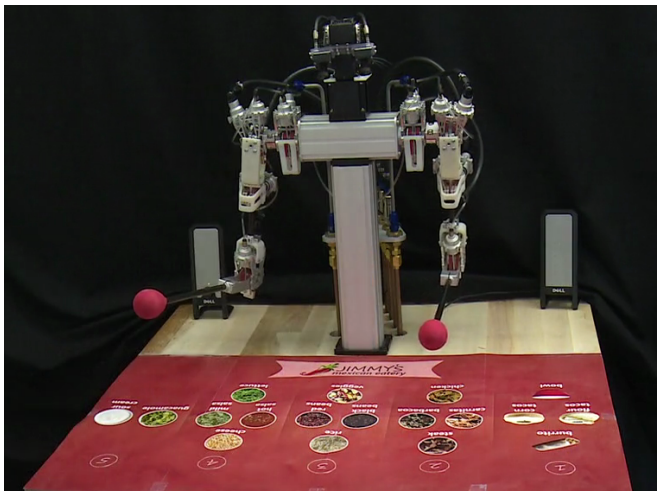

Fig. 3: Our robotic server and restaurant menu.

TABLE I: Sample dialogue from an Easy script. Red text inside square brackets contains the interface selections that the operator should perform at that moment.

Actor	**Hi.**
Operator	**[Hi]**
Jimmy	Welcome to Jimmy's. How can I get you started?
Actor	**Let's see. I'm going to have a burrito.**
Operator	**[Burrito], [Done]**
Jimmy	Burrito! Okay, and what meat would you like?
Actor	**Steak, please.**
Operator	**[Steak], [Done]**
Jimmy	Steak! Sure. Rice or beans or veggies?
Actor	**[silence]**
Operator	**[silence]**
Jimmy	Maybe some rice?
Actor	**Sure, rice is fine.**
Operator	**[Yes]**
Jimmy	Something else?
Actor	**Not from there.**
Operator	**[No]**
Jimmy	Okay, what about some toppings?
Actor	**Do you have jalapeños?**
Operator	**[RHM No]**
Jimmy	No.
Actor	**Okay, hot salsa then.**
Operator	**[More], [Hot Salsa], [Done]**
Jimmy	Hot salsa! Anything else?
Actor	**No, that's it.**
Operator	**[No]**
Jimmy	How about sour cream or guacamole to finish?
Actor	**Yeah, I'll have both**
Operator	**[Yes], [Both]**
Jimmy	Great. Here you go. Thanks for visiting Jimmy's.
Actor	**Bye!**
Operator	**[Bye]**
Jimmy	Bye-bye.

the use of the RHM, and silence). We define Level 2 items as items that require accessing an interface sub-menu to complete the order and Level 1 as those that do not. Hard scripts included 15 turns and were matched for all of the previously mentioned customer phenomena as well as four additional types (pointing gestures, multiple selections within the turn, a request for extra or double of an item, and a self-correction that might involve backing up in the menu). The Training script contained one instance of each type of customer behavior so that participants would have practice with a similar interaction.

C. Questionnaires

After participating in each condition, participants answered questions from The Temple Presence Inventory [17] and the Presence Questionnaire [18], giving ratings about the extent of mental immersion, the engagement of all of their senses, their ability to observe the customer's verbal and nonverbal behavior, the amount that the visual display quality interfered with or distracted from task performance, the degree to which the control devices interfered with task performance, and the degree to which the experience was emotional versus unemotional. After completion of the entire experiment, they answered an additional questionnaire asking which role they preferred, which roles they found easiest for nonverbal and verbal control, and how difficult they found it to coordinate their behaviors alone and in the pair.

D. Procedure

Upon arrival at the lab, participants completed consent forms and a demographic survey before receiving instructions about their roles during the study. Next, the experimenter described the purpose of the study and introduced the telepresence robot. Then participants were familiarized with the controllers and VR headset, and they learned how to use the AR interface. Finally, they were introduced to the restaurant scenario and task-specific instructions were given. Participants were asked to try to commit the fewest possible ordering/verbal selection mistakes and to coordinate their speech and gestures to the best of their abilities. To encourage the coordination of behaviors in every session, we highlighted the importance of tapping ingredients at the same time that speech was produced. Participants used the training script to practice in each of the four configurations until they achieved proficiency. We defined proficiency as performing the training script with no more than one verbal error and no more than one missed ingredient tap. In the training stage, we defined a maximum time of three seconds for the tap/verbal coordination to occur. When participants were training in the single operator condition, the other participant was asked to observe the robot from the customer's perspective to assess the importance of coordinating verbal and nonverbal behaviors.

The experimenter acted as the customer during the training stage but was replaced by a professional actor for the testing phase to reduce habituation and to ensure that the scripts were always performed in the same manner.

Each participant was assigned either the A or B role to determine his/her individual task order. Each session included four phases of training and testing: two single operator phases (Solo-A followed by Solo-B) and two paired phases (Pair Nonverbal A (PNV-A) and Pair Verbal B (PV-B) followed by PNV-B and PV-A). All training phases occurred before any of the testing phases. We counterbalanced the order in which the participants performed the single operator phase and paired-operator phases such that half of the participants used the single operator system first and the other half used the paired-operator system first. In single operator interactions, the actor used one Easy and one Hard script to test one participant's performance while the other participant waited outside of the room. In paired-operator interactions, the actor used the four remaining scripts to perform one Easy and one Hard script per pair condition. The experimenter instructed participants to immediately complete the 6-item presence questionnaires between each testing session.

E. Verbal and Nonverbal Analyses

In addition to the questionnaires, we used a set of six verbal and nonverbal performance metrics. These metrics, analyzed in different parts of the interaction, give us an objective insight into participants' performance.

1) Verbal Errors: Log files from the dialog system were compared to the scripts to identify selection errors in the LHM and calculate a Verbal error rate (the number of errors divided by the number of turns).

2) Overall Selection Response Time: Overall verbal response time calculations were performed by comparing the timestamp for the end of the customer's speech to the timestamp of the first subsequent menu selection, which includes any LHM or RHM choice. In some situations, the response occurred during one of the actor's lines. In these cases, the response time is calculated as zero. In all other cases, the difference between the end of the customer's vocalizations and the next menu choice is computed. An average value for all response times per session was used in analysis.

3) Ingredient Selection Response Time: We calculated the ingredient selection response time by extracting individual statistics specifically for ingredient selection (as opposed to the overall response time measures which include any selection on the interface). This measure includes ingredients at each level of menu and sub-menu. Those that need to be accessed using sub-menus may take significantly longer times to select.

4) Number of RHM Selections: Every script included a turn where participants were forced to use the RHM to continue the interaction. Additional usage of this menu was not necessary, but participants were free to use it at will for naturalness. This metric counts the number of times that the RHM was used excluding the forced point in the interaction.

5) Coordination Offset: The videos recorded from the actor's perspective were manually annotated to identify when the robot tapped a particular ingredient and which ingredient was selected. In addition to manual annotations of when and where taps occurred, automatic annotations were performed to extract the start and end times of each ingredient vocalization by the robot. For each ingredient selection annotation, if a tap occurred within the duration of the robot's vocalization, no offset was added to the session. If the tap occurred between the beginning and up to three seconds before the vocalization occurred, or between the end and up to three seconds after, we calculated a coordination offset that represented the absolute value of the time difference. We repeated this process for each ingredient in each session and calculated the coordination offset average.

6) Nonverbal Errors: If a verbalized ingredient was not matched with any tap within the three-second negative or positive interval described above, we counted it as a nonverbal error.

VI. Results

This section reports the results of three Restricted Maximum Likelihood (REML) [19] analyses. We examine the measures of verbal behavioral performance, explore nonverbal performance, and finish by presenting the evaluations from the questionnaires.

A. Verbal Performance

We performed an REML analysis to examine the effects of the manipulations on verbal errors, ingredient selection response times, overall selection response times across the various conditions, and use of the RHM (see Table II). Our independent variables are coded as follows: Solo and Pair denote whether the participant was performing the task alone or with a partner while controlling the verbal interface; Difficulty is whether the script being used by the actor was Easy or Hard; and Order signifies whether a participant did the task Solo or in a Pair first. In these analyses, only Pair Verbal data are considered of the Pair conditions.

1) Verbal Errors: We examined the error rate for all verbal output and found no significant effects of Solo/Pair, Difficulty, or the interactions between Order and Solo/Pair or Solo/Pair and Difficulty. There was a significant interaction between Order and Difficulty ($F = 6.259, p = 0.014$), and pairwise comparisons (alpha $= 0.05$) revealed significant differences between performing the Solo First and Hard combination of conditions relative to Solo First and Easy, Pair First and Easy, and Pair First and Hard. The combination of Solo First and Hard produced the most verbal errors. Additionally, there was a trend for Order ($F = 3.379, p = 0.077$) such that doing Solo First had a slightly higher error rate than doing Pair First.

2) Overall Selection Response Time: Across all trials, we found a main effect on overall verbal response time of Solo/Pair (Solo $= 1.474$, Pair $= 1.186$, $F = 26.450, p < 0.0001$) such that the verbal response time was slower in the Solo condition than in the Pair condition. Additionally, there was a main effect of Difficulty (Easy $= 1.467$, Hard $= 1.193$, $F = 21.960, p < 0.0001$), with Hard scripts having shorter average response times than Easy scripts.

TABLE II: Behavioral Performance–Verbal measures. Errors = Error rate, RT = Response Time, RHM = Right-hand menu usage.

	Errors	Overall RT	Ingred. RT	RHM
Solo Easy	0.08	1.58	2.77	0.20
Solo Hard	0.11	1.37	4.79	0.73
Pair Easy	0.07	1.36	2.08	0.23
Pair Hard	0.07	1.11	3.51	0.27

TABLE III: Behavioral Performance–Nonverbal measures. CT = Coordination time, CE = Coordination Errors, S = Simple, M = Multi, A = Average.

	CT-S	CT-M	CT-A	CE-S	CE-M	CE-A
Solo Easy	0.30	0.30	0.35	0.10	0.10	0.20
Solo Hard	0.30	0.34	0.32	0.03	0.57	0.60
Pair Easy	0.32	0.42	0.35	0.03	0.10	0.13
Pair Hard	0.38	0.52	0.48	0.03	0.37	0.40

3) Ingredient Selection Response Time: For selecting ingredients specifically, we identified a significant effect of Solo/Pair (Solo $= 3.772$, Pair $= 2.795$, $F = 60.061, p < 0.0001$) such that the time for Solo performance was much longer than for Pair performance. There was also a significant effect of Difficulty (Easy $= 2.413$, Hard $= 4.156$, $F = 188.992, p < 0.0001$) wherein the verbal response time was longer for the Hard scripts than for the Easy scripts. There was a significant interaction between Solo/Pair and Difficulty ($F = 6.353, p = 0.014$) such that Solo Hard was significantly slower than Solo Easy or Pair Hard, which were in turn slower than Pair Easy. There were no significant effects of Order or other interactions.

To determine whether complexity affected verbal response time, we analyzed subsets of these data based on the number of menu levels through which the participant had to navigate to complete the turn. For the simple Level 1 selections (i.e., those that did not require changing menu levels), there were still significant effects of Solo/Pair (Solo $= 2.534$, Pair $= 2.188$, $F = 7.763, p = 0.007$) and Difficulty (Easy $= 1.869$, Hard $= 2.858$, $F = 56.970, p < 0.0001$), with no significant effects of order or interactions. For the multi-level (Level 2) selections, we found a similar pattern in which Solo/Pair (Solo $= 4.806$, Pair $= 3.460$, $F = 40.983, p < 0.0001$) and Difficulty (Easy $= 3.584$, Hard $= 4.682$, $F = 28.340, p < 0.0001$) had significant effects, but there were no effects of order or the interactions.

4) RHM Usage: In total, there were 28 Solo and 15 Pair usages of our interface's RHM when it was not directly required by the script. There was a significant main effect of Difficulty such that this menu was used more during Hard scripts (Easy $= 0.223$, Hard $= 0.511$, $F = 5.850, p = 0.018$) and a trend towards an effect of Solo/Pair such that the RHM was used more during the Solo condition (Solo $= 0.480$, Pair $= 0.254$, $F = 3.586, p = 0.062$). The interaction between Difficulty and Solo/Pair was also significant ($F = 4.812, p = 0.031$), with pairwise comparisons indicating that participants used the RHM significantly more in the Solo Hard condition relative to all other conditions.

B. Nonverbal Coordination

To assess nonverbal coordination, we contextualized nonverbal behaviors with the verbal output as described in the evaluation section. We performed REML analysis to examine the effects of our manipulations on coordination offsets and nonverbal errors, this time considering Solo and Pair Nonverbal data. (See Table III.)

1) Coordination Offsets: For the coordination offsets across all trials, there was a trend towards a main effect of Solo/Pair (Solo $= 0.330$, Pair $= 0.414$, $F = 3.162, p = 0.079$). There were no significant effects of Solo/Pair, Difficulty, or Order on the the simple (Level 1) selection coordination time. For the multi-level (Level 2) selection coordination time, there was a significant effect of Solo/Pair where Pair offsets were larger (Solo $= 0.318$, Pair $= 0.472$, $F = 4.649, p = 0.034$).

2) Nonverbal Errors: The number of coordination errors for simple (Level 1) selections was unaffected by Solo/Pair, Difficulty, and Order manipulations. However, there was a significant main effect of Difficulty for multi-level (Level 2) selections (Easy $= 0.100$, Hard $= 0.467$, $F = 13.032, p = 0.0005$) and total coordination errors across both selection types (Easy $= 0.167$, Hard $= 0.500$, $F = 7.375, p = 0.008$).

C. Questionnaires

We performed REML analysis on the six questions that participants answered after performing the task in each condition to examine the effects of condition and order (see Figure 4). We found a significant effect of condition on the extent of mental immersion ($F = 16.879$, $p < 0.0001$), and pairwise comparisons revealed that the responses for the Solo and Pair-Nonverbal Operator (PNV) conditions were significantly higher than for the Pair-Verbal Operator (PV) condition. There was no effect of order. We found a similar pattern of effects for ratings of how completely all senses were engaged ($F = 35.647$, $p < 0.0001$, Solo $>$ PV, PNV $>$ PV), how high the participant rated his/her ability to observe the actor's verbal and nonverbal behaviors ($F = 6.191$, $p = 0.004$, Solo $>$ PV, PNV $>$ PV), and how emotional the participants rated the experience ($F = 6.560$, $p = 0.003$, Solo $>$ PV, PNV $>$ PV). Again, there were no effects of order. Additionally, we found no significant condition or order effect for ratings of how much the visual display quality interfered with or distracted from task performance. Finally, we found a significant effect of condition for how much the control devices interfered with task performance ($F = 6.057$, $p = 0.004$, Solo $>$ PNV, PV $>$ PNV) such that PNV had the lowest interference ratings. There was also an effect of order on this question in which performing the Solo conditions first resulted in higher ratings than performing the Pair conditions first ($F = 6.413$, $p = 0.017$, Solo $>$ Pair).

For the final questionnaire, we performed Chi Squared tests to analyze participant preferences. There was a significant difference in the answers to "Which was your favorite role?" ($\chi^2 = 9.437, p = 0.009$) such that the highest number

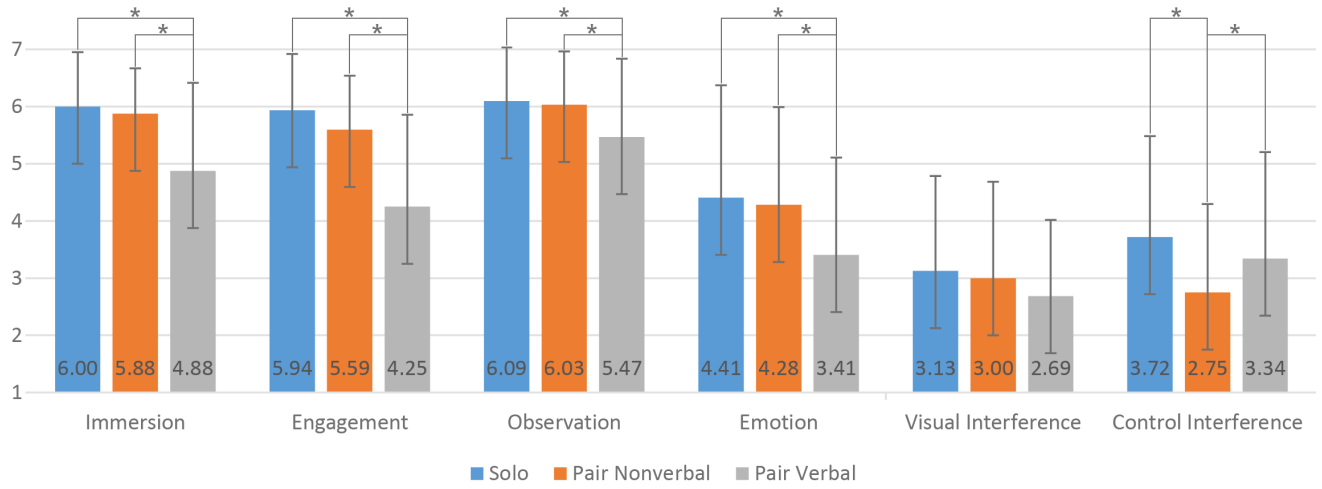

Fig. 4: Presence Questionnaire Responses. Ratings were provided on a 7-point scale.

TABLE IV: Preference Results

Question	Solo	PNV	PV	Prob.
Prefer	17	12	3	0.009
Easiest NV	10	22	N/A	0.034
Easiest V	14	N/A	18	0.480

of participants prefered the Solo condition, followed by the PNV and PV conditions. (See Table IV.) There was also a significant difference in responses to "In which role did you find it easiest to control the robot?" ($\chi^2 = 4.500, p = 0.034$) such that participants found PNV easier than Solo. Finally, there was no significant difference in responses to the question, "In which role did you prefer to control the verbal interface?" ($\chi^2 = 0.500, p = 0.480$), such that PV was similar to Solo.

VII. Discussion

As expected, Hard scripts elicited more verbal errors than Easy ones. However, contrary to our expectations, the results showed no significant effects on verbal error rates of performing the task alone or as a pair. The results also showed a trend for an effect of order. Participants that first interacted in the Solo condition performed about 71% of the total Solo verbal errors. These results can be explained by the added difficulty of learning while simultaneously controlling the two different modalities in contrast to learning to use one subcomponent of the system at a time. After a steeper learning curve in the Solo condition, error rate performance seemed to stabilize. To further explore this finding, we examined verbal response times. Both overall and ingredient response times are significantly higher in the Solo condition. It seems that to maintain a low error rate, some participants take their time to respond when interacting in the Solo condition. The higher response times can be attributed to the higher cognitive load introduced by simultaneously controlling the nonverbal modality, to the less ergonomic joysticks placed on the robot's arms, or both. Participants that interacted first in the Solo condition also took longer than those who started as a pair. These results again suggest a deeper learning curve for the Solo condition or a preferable training order. The ingredient response time for the Hard scripts (> 3.5 seconds) shows the difficulty of navigating our menus to make complex patterns of selections. The delay increases more than 1.2 seconds when comparing ingredient response time in Solo Hard versus Pair Hard. Focused analyses confirmed that menu navigation affects response time by showing that having to use a Level 2 menu results in longer times than using a Level 1 menu.

Although there were no significant effects of the various manipulations on the coordination time on average or for the simple (Level 1) selections alone, there was a significant effect of Solo/Pair for the multi-level (Level 2) selections. Pairs of operators had larger coordination offsets than single operators for these complex selections, indicating that it is harder to coordinate across two operators than act alone when navigating menus becomes more difficult. Nonverbal errors where the operator did not tap the ingredient were not common given the emphasis the experimenter put on this behavior in the training sessions. However, some errors still occurred, and those few errors were particularly likely to occur when navigating a multi-level menu selection in a hard interaction. Anecdotal observations from the experimenters suggest that other performance features such as gazing at the participant at the right time when speaking or complementing an RHM option, such as a "No", with the shake of the head were more common and coordinated in the Solo condition. However, some post-experiment comments suggest that pair communication protocols established over time could help to alleviate these effects. Future work is needed to examine these issues further.

Responses to presence items on the questionnaire are not significantly different when we compare solo operators with nonverbal operators in a pair, but both are rated significantly higher when compared with verbal operators in a pair. We

found that controlling the robot's nonverbal behaviors either alone or as a member of a pair elicited higher ratings of mental immersion, sensory engagement, ability to observe the conversation partner, and amount of emotion relative to controlling verbal behaviors in a pair. The added sense of presence in the solo interaction is arguably more important because it might guide the operator to make better choices and use the interface differently given the improved ability to observe the conversation partner. If the operator is solely controlling the nonverbal modality and not the verbal interface, he or she can't use the improved engagement and immersion in the environment to react to important cues. We observe this behavior in our data as differences in the usage of the RHM. This difference is due to the increased utilization that Solo participants make of options like "OK" to reply to the actor's requests and "Sorry" when they commit a mistake. These behaviors were not necessary and not rehearsed during the training session, and we speculate that their frequency increase is due to the higher sense of presence in the environment.

VIII. Conclusion

We have presented and evaluated an AR dialog interface that enables multimodal teleoperation of robots. According to the insights gathered from a user study, opting to use two operators instead of one may make sense in scenarios where fast response times are required and/or in situations where the operators have little need to coordinate verbal and nonverbal behavior. However, there remains the inherent problem of the increased cost of having to train and maintain two different operators and manage problems that arise from the shared communication that has to be established and followed. By using AR/VR tracking technology, a manipulable input device and an interface designed to reduce cognitive load while the operator is performing, we can successfully eliminate the need for a second operator without sacrificing accuracy of task performance. Most of the negative outcomes found in the Solo condition suggest an order effect that could be resolved with a part-whole training design or more extended training phase. Our results also suggest that using our system to control the robot results in an increased sense of presence and, non-trivially, the single-operator setup was preferred by our participants.

Presence is defined as the subjective experience of being in one place or environment, even when physically situated in another [18]. In our context, an added sense of presence is important because it leads to a better ability to observe and understand the other members of the interaction. We argue that being able to control the gaze, gestures, and verbal behaviors of a robot while benefiting from an increased sense of presence is a powerful tool to control telepresence robots, to collect verbal interaction data to train multimodal autonomous systems and to conduct WoZ interactions. The AR interface presented here is useful in any scenario where an operator has to use full body motions to control a robot's gestures and gaze while at the same time verbally interact with users without being able to use his or her own voice.

References

[1] I. Leite and J. F. Lehman, "The robot who knew too much: Toward understanding the privacy/personalization trade-off in child-robot conversation," in *Proceedings of the The 15th International Conference on Interaction Design and Children*. ACM, 2016, pp. 379–387.

[2] S. Kasahara, M. Ando, K. Suganuma, and J. Rekimoto, "Parallel eyes: Exploring human capability and behaviors with paralleled first person view sharing," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. ACM, 2016, pp. 1561–1572.

[3] S. Tregillus and E. Folmer, "Vr-step: Walking-in-place using inertial sensing for hands free navigation in mobile vr environments," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI '16. ACM, 2016, pp. 1250–1255.

[4] H. Martins and R. Ventura, "Immersive 3-d teleoperation of a search and rescue robot using a head-mounted display." in *ETFA*, 2009, pp. 1–8.

[5] K. Higuchi and J. Rekimoto, "Flying head: A head motion synchronization mechanism for unmanned aerial vehicle control," in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '13. ACM, 2013, pp. 2029–2038.

[6] C. Pittman and J. J. LaViola, Jr., "Exploring head tracked head mounted displays for first person robot teleoperation," in *Proceedings of the 19th International Conference on Intelligent User Interfaces*, ser. IUI '14. ACM, 2014, pp. 323–328.

[7] S. Kratz, J. Vaughan, R. Mizutani, and D. Kimber, "Evaluating stereoscopic video with head tracking for immersive teleoperation of mobile telepresence robots," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, ser. HRI'15 Extended Abstracts. ACM, 2015, pp. 43–44.

[8] L. Fritsche, F. Unverzag, J. Peters, and R. Calandra, "First-person tele-operation of a humanoid robot," in *Humanoid Robots (Humanoids), 2015 IEEE-RAS 15th International Conference on*. IEEE, 2015, pp. 997–1002.

[9] K. Zheng, D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "How many social robots can one operator control?" in *Proceedings of the 6th international conference on Human-robot interaction*. ACM, 2011, pp. 379–386.

[10] P. Baxter, J. Kennedy, E. Senft, S. Lemaignan, and T. Belpaeme, "From characterising three years of hri to methodology and reporting recommendations," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2016, pp. 391–398.

[11] L. D. Riek, "Wizard of oz studies in hri: a systematic review and new reporting guidelines," *Journal of Human-Robot Interaction*, vol. 1, no. 1, 2012.

[12] D. V. Lu and W. D. Smart, "Polonius: A wizard of oz interface for hri experiments," in *Proceeding of the 6th ACM/IEEE international conference on Human-robot interaction*, ser. HRI '11. ACM, 2011.

[13] M. Villano, C. R. Crowell, K. Wier, K. Tang, B. Thomas, N. Shea, L. M. Schmitt, and J. J. Diehl, "Domer: A wizard of oz interface for using interactive robots to scaffold social skills for children with autism spectrum disorders," in *Proceedings of the 6th International Conference on Human-robot Interaction*, ser. HRI '11. ACM, 2011, pp. 279–280.

[14] G. Hoffman, "Openwoz: A runtime-configurable wizard-of-oz framework for human-robot interaction," in *2016 AAAI Spring Symposium Series*, 2016.

[15] J. P. Whitney, T. Chen, J. Mars, and J. K. Hodgins, "A hybrid hydrostatic transmission and human-safe haptic telepresence robot," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 690–695.

[16] S. R. Fussell, R. E. Kraut, and J. Siegel, "Coordination of communication: Effects of shared visual context on collaborative work," in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 2000, pp. 21–30.

[17] M. Lombard, T. B. Ditton, and L. Weinstein, "Measuring presence: the temple presence inventory," in *Proceedings of the 12th Annual International Workshop on Presence*, 2009, pp. 1–15.

[18] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence: Teleoperators and virtual environments*, vol. 7, no. 3, pp. 225–240, 1998.

[19] M. G. Kenward and J. H. Roger, "Small sample inference for fixed effects from restricted maximum likelihood," *Biometrics*, pp. 983–997, 1997.