Multimodal Metric Study for Human-Robot Collaboration

Scott A. Green
s.a.green@lmco.com

Scott M. Richardson
scott.m.richardson@lmco.com

Randy J. Stiles
randy.stiles@lmco.com

Lockheed Martin Space Systems Company, Advanced Technology Center

Mark Billinghurst
Human Interface Technology Laboratory,
New Zealand, HIT Lab NZ
mark.billinghurst@canterbury.ac.nz

J. Geoffrey Chase
Mechanical Engineering Department,
University of Canterbury, New Zealand
geoff.chase@canterbury.ac.nz

Abstract

The aim of our research is to create a system whereby human members of a team can collaborate in a natural way with robots. In this paper we describe a Wizard of Oz (WOZ) study conducted to find the natural speech and gestures people would use when interacting with a mobile robot as a team member. Results of the study show that in the beginning participants used simple speech, but once the users learned that the system understood more complicated speech, they began to use more spatially descriptive language. User responses indicate that gestures aided in spatial communication. The input mode that combined the use of speech and gestures was found to be best. We first discuss previous work and detail how our study contributes to this body of knowledge. Then we describe the design of our WOZ study and discuss the results and issues encountered during the completion of the experiment.

1. Introduction

The design of interfaces for Human-Robot Interaction (HRI) will be one of the greatest challenges that the field of robotics faces [1]. It's obvious that if robots and humans are going to become collaborative partners, appropriate interfaces must be created to enable Human-Robot Collaboration (HRC).

We are developing an interface that enables humans to collaborate with robots through the use of natural speech and gesture combined with a deeper understanding of spatial context and a rich spatial vocabulary. We have a current working prototype of the Spatial Dialog System (SDS) [2], but need to determine what type of speech and gestures a human team member would use to collaborate with a robotic system. With this in mind, we have designed a Wizard of Oz (WOZ) study to determine the type of speech and gestures that would be used. The results of this study will be used to enhance the development of our current spatial dialog system.

2. Related Work

Bolt's work "Put-That-There" [3] showed that gestures combined with natural speech (multimodal interaction) lead to a powerful and more natural man machine interface. We conducted a WOZ study to enable the development of robust multimodal interaction for our SDS platform. A WOZ study is one where the system is not fully functional and a human wizard acts for the parts of the system that have not yet been implemented. The participants in a WOZ study do not know that a human is involved; they are instructed to interact with the system as if it were fully operational.

For example, Makela *et al.* [4] found their WOZ study to be instrumental in the iterative development of the Doorman system. The Doorman is used to control the access of visitors and staff to their building and also to guide visitors upon entry into the building. Their study was conducted where the human wizard completed speech recognition and the rest of the system was operating normally. From their study they found that they needed to shorten the utterances from the system to reduce communication time, to provide the user with feedback to confirm that the system is operational, and have better error handling.

To find out what kind of speech would be used with a robot in grasping tasks, Ralph *et al.* [5] conducted a user study whereby users were asked to tell a robot to pick up five different small household objects. The robot was fixed on a table and the users sat next to the robot when giving it instructions. The participants were asked to be as descriptive as possible in their commands and a human operator translated these commands into robot movement. Participants felt that natural language was an easy way to communicate with the robotic system and all participants were able to complete the pick and place tasks given to them. Participants did tend to use short commands in a mechanical manner.

A WOZ experiment was used by Carbini *et al.* [6] for a collaborative multimodal story telling task. The objective of the study was to determine what speech and gestures would be used as two participants collaborated remotely with the system to create a story. In this study the human wizard completed the commands of the users' speech and gestures from a laser pointer. Users in this study were found to complete a laser pointing gesture with an oral command, and the users tended to point without stretching their arms.

A similar study to the one we have conducted is by Perzanowski *et al.* [7]. They, too, are designing an intuitive way to interact with intelligent robotic systems in a multimodal manner. In their pilot WOZ study they focused on verbal communication and gestural input through a touch screen to collaborate with a remotely located mobile robot. Perzanowski *et al.* [7] were interested in finding out how people referred to objects when giving directions and trying to maneuver a mobile robot. Participants were told they could talk to the robot as if it were human and could point to objects and locations on a touch screen that included ego- and exo-centric viewpoints. The participants were told to get the robot to find an object. Two wizards interpreted the speech and touch gestures and drove the robot where they interpreted the user wanted the robot to go and spoke for the robotic system. Users felt they had to continually guide the robot and so used a lot of short spoken commands. If the users had felt the robot was more autonomous they may have used more complex speech.

Our study is novel in that the participants were able to use speech and free hand natural gestures to control a mobile robot. The study by Perzanowski *et al.* [7] allowed for full use of speech, but the gestures used were constrained to those of pointing at a touch screen. The objective of our study was to find out what combination of speech and free hand natural gestures would be used when collaborating with a mobile robot on a navigation task. Unlike previous studies we also split the modalities and ran a test for speech only, gesture only and speech and gesture combined. In this manner we can compare how users changed their interaction with the mobile robotic system based on what modality was available to them.

3. Wizard of Oz Study

Participants guided the robot through a maze and were told that the robot was autonomous but that its sensors had failed, i.e. it could not see. The participants had an exo-centric view of the maze and robot in addition a view from the camera mounted on the robot; see Fig. 1. Thus, the objective for the participants was to work with the robot and guide it through the maze using combined speech and gestures. Users were told that the system was practically fluent in understanding spatial dialog and gestures including combined speech and gesture input.

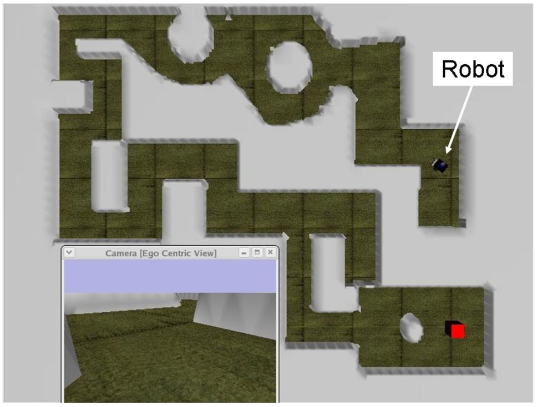

Fig. 1: Example of maze for participants to guide robot through using speech and gestures.

A pre-experiment questionnaire was given to each participant to find out what type of speech and gestures they would like to use. Subjects were asked what speech, gestures and speech combined with gestures they would use with a human collaborator. The user was shown pictures illustrating where the human collaborator was to move, from point A to point B.

Pictures were used so that the participants would not be biased with spatial language that would have been contained in a written or verbal question. The view of the pictures in the questionnaire was varied to test what reference frames the participants would use. If the reference frame of the robot was not aligned with the user then we wanted to see what spatial references would be used.

To determine if participants would communicate differently with a robot as opposed to a human a similar questionnaire was given out after the experiment was run. This questionnaire was similar to the pre-experiment questionnaire except that instead of the human, the participants were questioned about how they would guide a robot from A to B.

One question was repeated for each modality for both the human and robot cases. One time the picture indicated to go around an unidentifiable object. Another time the picture indicated to go around a pizza, something most participants could identify with. Fig. 2 shows both the unidentifiable and identifiable objects. The point of these questions was to see if the user would indicate to go around "this" or "around the pizza".

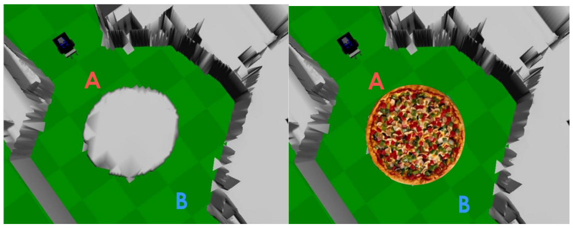

Fig. 2: Question indicating robot to go around unidentifiable object (left) and around the pizza (right).

After the pre-experiment questionnaire was completed, the participants were told that they would be working with a robot lunar rover. The rover had experienced sensor failures and they were to collaborate with the robot and get it back to safety. The robot was simulated using Gazebo [8] from the Player/Stage project. The video output from Gazebo was projected onto a screen that the user stood in front of. Cameras were placed so that the gestures used by the participant could be seen by the system. A microphone in the ceiling picked up the user's voice. The users were told that the system was capable of understanding most verbal and gestural spatial references and that they should use a wide variety of speech and gestures.

Unknown to the users a wizard was observing their speech and gestures and driving the robot accordingly. The same wizard was used for all participants to reduce the chance of varying interpretations of the participant's speech and gesture. The wizard responded to the user if speech or gestures were used that were not understood with canned responses selected by keyboard input. The wizard also used canned responses to alert the user when the experiment would begin, what modality would be used, when they had reached the goal position and if they had crashed into a wall.

Each participant collaborated with the robot to go through the maze three separate times. The maze had multiple curves and forks so the user would have to use a variety of spatial language. The participants had both an exo (God-like) and ego (robot's view) of the workspace.

Three conditions used were:

- Speech only: users were told to only use speech command
- Gesture only: users were told to only use gesture commands
- Combined speech and gesture: users were told they could use a free mixture of speech and gesture

A post-experiment questionnaire was given to the participants with answers provided on a Likert scale of 1-7 (1 = strongly agree, 7 = strongly disagree). The questions intended to gauge user satisfaction with the system and modality preference. Post experiment interviews helped to determine whether the participants felt the system was actually operational and not driven by a wizard.

4. Results

We ran the study with 10 participants recruited from within Lockheed Martin Space Systems Company, Advanced Technology Center. The group consisted of nine engineers and one person from Finance. There was one female and nine males all under the age of 25. The responses to the demographic questionnaire showed that overall the group was not familiar with either robotic systems or speech systems and claimed they generally used gestures when speaking.

4.1 Pre-Experiment Questionnaire

1) Speech Only

When guiding the person from point A to point B for left and right turns users primarily used the term "turn" (9 right and 9 left), while one used rotate (right) and one used references to a clock, i.e. 7 o'clock then 4 o'clock. Three participants included an angle with the command "turn", such as "turn right 90 degrees". To indicate forward movement users used a combination of the following commands: move, go, forward, straight and walk.

For the case of moving around the unidentifiable object and pizza, the participants used the same commands for both cases. This is not what we expected, we thought the users would use go around "this" for the unidentifiable case, but no one did. Eight participants gave incremental instructions, such as forward, stop, turn right 45 degrees, stop, forward, turn left 45 degrees, stop, forward, turn left 45 degrees, stop, turn right 45 degrees, stop, forward, stop. Two participants used the preposition around and identified the pizza to go around.

2) Gesture Only

Participants indicated they would use finger gestures (5) or full arm gestures (4) with the remaining user having a preference to use arm gestures analogous to those for riding a bike. Right and left turns were instructed with either a full arm out in the appropriate direction or a similar instruction using only fingers. One participant indicated pointing to relative locations on a clock.

The gesture for stop was fairly consistent for all users. Hands up with palm out indicated stop. One user used a quick up and down motion of the fingers to indicate stop with one user using a fist to indicate stop.

3) Speech Combined with Gestures

Participants combined the answers for the speech-only case and gesture-only case for the combined speech and gesture questions. Typically the answers had the speech from the speech-only case complemented with the answers from the gesture-only case. This result is likely due to the users not wanting to repeat themselves so used common answers, i.e. answers already developed, instead of answering the questions from the very beginning.

4) Comparison to Questionnaire with Robot

A similar questionnaire to the pre-experiment one was given out after the study with the person replaced by the robot from the experiment. The intent of this questionnaire was to see how the user's responses changed after running the experiment and to see if the communication with the robot differed greatly than that with a person.

The communication indeed became more mechanized for the case with the robot. Each step was given incrementally with turns provided as discrete angles, except for one user who instructed the robot to "turn around the corner". The communication to the robot was simple, short and curt such as move, turn, and stop type utterances.

4.2 Experimental Results

1) Speech

Participants tended to use the same verbal references for stop and turn as reported in the questionnaire. Stop was simply "stop, for turning they used "turn" and "rotate". Magnitudes were sometimes associated with the turn and rotate commands whilst some of the participants followed a turn command with a stop command. New terms were used to indicate the robot move forward, such as "walk", "drive" and "inch forward". Users at times were required to have the robot move backwards, for this they used the two terms "backwards" and "reverse".

An interesting result was the type of modifiers used. For example, to correct the robot when it had turned too far, users would say "back to the left". If the robot had not rotated the amount the user expected, this was corrected with phrases such as "a little bit more", "until I say stop" and "some more".

Participants spoke in mechanized terms when they first started the experiment, as experienced by Perzanowski *et al.* [7]. If something unexpected happened, like a crash was impending, then the users would resort to communicating with the robot like it was a team member and not as if it were a robot. Once users felt comfortable with the system and its capabilities, they began to use more descriptive speech than just "turn", "move" and "stop".

Users commented after the experiment "once I started using more complicated instructions than simple 'go forward' and 'turn' it became easier to control". An example of this type of interaction was when one user kept the robot moving forward and would tell it to turn around the corners without stopping forward movement. Through the second half of the maze for this run robot movement was much smoother, as opposed to the turn, stop, move, stop commands given in the first half of the maze.

2) Gesture

To have the robot move forward most users held their hand out at arms length in front of them. One user held the index finger up and then brought it down toward the screen in front of them to indicate move forward. Most users gave a gesture for the robot to move and then released the gesture. One user, however, maintained gestures the entire time the move was desired, i.e. the entire time the robot was to move forward the participant would keep his arm stretched out in front of him. Naturally, afterwards the user commented on how tired his arms were at the end of the trial.

The gesture for stop was consistent between all users. Hands ups, whether directly in front of the body or at full arms length, with palm towards the camera. One or two hands were used; this varied between users and also varied within the same trial of individual users.

Gestures for turning consisted of a full arm gesture to the side of the body that the user wanted the robot to turn in. All participants used the reference frame of the robot. Three users adjusted the degree of the turn by starting with the forward gesture (arm extended out in front of them) and defining the turn by how far their arm moved to one side.

3) Speech Combined with Gestures

Participants tended to use the same methodology for guiding the robot in the multimodal mode as in the speech only and gesture only modes. This methodology for seven participants consisted of combining the techniques used in the verbal only and gesture only trials to guide the robot, but doing so in incremental steps go, stop, turn, stop, go etc.

Three participants used more complex communication, such as “go around this’ whilst using a full arm gesture to indicate a turn, or “go around the corner to your right”, again whilst gesturing using a full arm extended to the side indicating to turn. The result of this type of communication was more fluid motion of the robot. When more descriptive communication was used, there were fewer stops for the robot which resulted in decreased time to complete the task.

The three participants who used the more descriptive communication that resulted in fewer stops all had completion times far less than the average. The average completion time for the multimodal case was 438.5 seconds; the three users with fluid robot motion had completion times of 272, 291 and 298 seconds. This result shows that using more complex communication enabled fluid robot motion that decreased completion times.

4) Times, Distances and Crashes

Although the multimodal case had the lowest average completion time, there was no significant difference in the time it took to complete the trials between the three modalities (ANOVA: F(2,24) = 1.73, p > .05). See Fig. 3 for average completion times. These three measures were dependent on the user and not the modality of communication. If a participant crashed in one modality, then the user tended to crash in all three. The distance traveled was also dependent on the user and not the modality as no significant difference was found between modalities (ANOVA: F(2,25) = 0.23, p>.05).

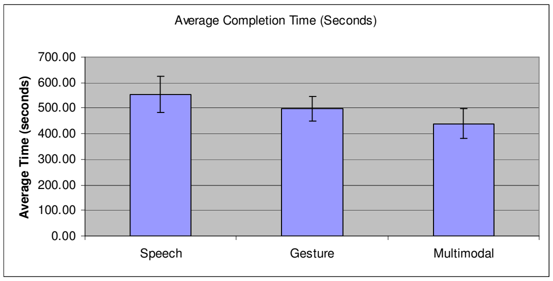

Fig. 3: Average completion times for the three modalities used.

4.3 Post Experiment Questionnaire

The post experiment questionnaire showed that users felt the system understood verbal spatial references very well, as should be the case since the wizard was interpreting speech. Participants felt that the system understood their gestures, although not as well as speech. This result is expected since the wizard was able to fully comprehend the speech used but had to interpret gestures, which took more time and when gestures were ambiguous the wizard didn’t always pick up on them.

Participants felt that the use of gestures helped them to communicate spatially with the system. Participants had high confidence speaking to the system and were relatively confident gesturing to the system. Participants felt that the multimodal (speech and gesture) mode was the best (ANOVA: F(2,27) = 4.09, p < .05) and that the gesture mode only was the worst, see Fig. 4.

5. Discussion / Design Guidelines

The goal of our study was to find out what kind of speech and gestures people would use to interact with a mobile robot. Users were encouraged not to repeatedly provide the same communication once they found out a given command worked, but to try new commands to see if the system would understand them. Given the opportunity participants used natural speech and gestures to work with a robotic team member. Initially participants communicated with the robot using short mechanized terminology (rotate, stop, forward, stop, etc.). However once the participants learned they

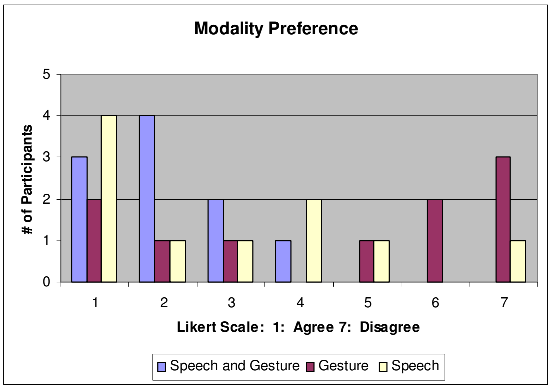

Fig. 4: User modality preference, users preferred the combined speech and gesture modality.

could communicate in a natural fashion they did so (go around that corner in front of you) and commented on the natural and intuitive nature of the interface.

Users preferred full arm gestures to indicate forward and turning motions. The system should react by initiating a turn and continuing to do so until a command is received to stop. One comment was made that the user preferred speech because then their arms “would not get tired”, so it’s important to think about ergonomics when designing gestures into a system.

A gesture for turning should also define the magnitude of the turn. A participant used one arm forward to indicate move forward and then used the other arm to continually make turns. When the turn would go from right to left, the user would change which arm was used for the forward motion (always maintaining this forward motion) and use the appropriate arm for gesturing a turn and its magnitude.

One participant commented that it would have been nice to interact with the visuals. The user would have liked to been able to touch a point on the screen and tell the robot to go “there”. This is encouraging news for our research as that is exactly what kind of interface we are working towards [2], using Augmented Reality as a means for enabling a user to pick out a point in 3D space and referring to it as “here” or “there”.

Interestingly, all users thought they were interacting with a functioning system. No one suspected that there was a wizard interpreting all verbal and gestural communication. We can only hope that this resulted in the users feeling comfortable with using natural speech and gestures during the experiment.

6. Conclusions

In this paper we described a Wizard of Oz (WOZ) study for Human-Robot Interaction (HRI) that we conducted. The next step in our research is to incorporate the results from this WOZ study into our current architecture. It is clear that given the opportunity, users prefer natural speech and gesture, so this type of communication will be incorporated into our system.

7. Acknowledgements

This work was supported by Internal Research and Development funds provided by Lockheed Martin Advanced Technology Center who we wish to thank for permission to publish this work.

8. References

[1] S. Thrun, "Toward a Framework for Human-Robot Interaction," *Human-Computer Interaction*, vol. 19, pp. 9-24, 2004.

[2] S. Green, S. Richardson, V. Slavin, and R. Stiles, "Spatial Dialog for Space System Autonomy," *In Proceedings of ACM/IEEE International Conference on Human-Robot Interaction*, pp. 341-348, 2007.

[3] R. A. Bolt, "Put-That-There: Voice and Gesture at the Graphics Interface," *In Proceedings of the International Conference on Computer Graphics and Interactive Techniques*, vol. 14, pp. 262-270, 1980.

[4] K. Makela, E. P. Salonen, M. Turunen, J. Hakulinen, and R. Raisamo, "Conducting a Wizard of Oz Experiment on a Ubiquitous Computing System Doorman," *In Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue*, pp. 115-119, 2001.

[5] M. Ralph and M. Moussa, "Human-Robot Interaction for Robotic Grasping: A Pilot Study," 2005.

[6] S. Carbini, L. Delphin-Poulat, L. Perron, and V. J. E., "From a Wizard of Oz Experiment to a Real Time Speech and Gesture Multimodal Interface," *Signal Processing Journal, Special Issue on Multimodal Interfaces, Elsevier*, 2006.

[7] D. Perzanowski, D. Brock, W. Adams, M. Bugajska, A. C. Schultz, J. G. Trafton, S. Blisard, and M. Skubic, "Finding the FOO: A Pilot Study for a Multimodal Interface," 2003.

[8] Player-Project, *http://playerstage.sourceforge.net*, 2007.