Robotica (2007) volume 25, pp. 521–527. © 2007 Cambridge University Press
doi:10.1017/S0263574707003414 Printed in the United Kingdom

Human–robot interaction via voice-controllable intelligent user interface

Harsha Medicherla† and Ali Sekmen‡*

†*Department of Electrical and Computer Engineering, Tennessee State University, 3500 John A. Merritt Blvd. Nashville, TN 37209, USA*
‡*Department of Computer Science, Tennessee State University, 3500 John A. Merritt Blvd. Nashville, TN 37209, USA*

(Received in Final Form: January 23, 2007. First published online: March 6, 2007)

SUMMARY
An understanding of how humans and robots can successfully interact to accomplish specific tasks is crucial in creating more sophisticated robots that may eventually become an integral part of human societies. A social robot needs to be able to learn the preferences and capabilities of the people with whom it interacts so that it can adapt its behaviors for more efficient and friendly interaction. Advances in human–computer interaction technologies have been widely used in improving human–robot interaction (HRI). It is now possible to interact with robots via natural communication means such as speech. In this paper, an innovative approach for HRI via voice-controllable intelligent user interfaces is described. The design and implementation of such interfaces are described. The traditional approaches for human–robot user interface design are explained and the advantages of the proposed approach are presented. The designed intelligent user interface, which learns user preferences and capabilities in time, can be controlled with voice. The system was successfully implemented and tested on a Pioneer 3-AT mobile robot. 20 participants, who were assessed on spatial reasoning ability, directed the robot in spatial navigation tasks to evaluate the effectiveness of the voice control in HRI. Time to complete the task, number of steps, and errors were collected. Results indicated that spatial reasoning ability and voice-control were reliable predictors of efficiency of robot teleoperation. 75% of the subjects with high spatial reasoning ability preferred using voice-control over manual control. The effect of spatial reasoning ability in teleoperation with voice-control was lower compared to that of manual control.

KEYWORDS: Human–robot interaction; Mobile robots; Speech recognition; Intelligent user interfaces.

1. Introduction

One of the overarching goals of robotics research is that robots ultimately coexist with people in human societies as an integral part of them. In order to achieve this goal, robots need to be accepted by people as natural partners within the society. It is therefore essential for robots to have human-like perception and interaction capabilities that can be utilized for effective human–robot interaction (HRI).

*Corresponding author. E-mail: asekmen@tnstate.edu

A social robot is defined as "*an autonomous or semi-autonomous robot that interacts and communicates with humans by following the behavioral norms expected by the people with whom the robot is intended to interact*."[1] A social robot needs to be able to learn the preferences and capabilities of the people with whom it interacts so that it can adapt its behaviors for more efficient and friendly interaction. Social Robotics focuses on the development of robots that operate with people to meet or address some social needs.[2] One active area of research in Social Robotics is investigating specifically how to socially equip robots to respond to the needs of the people. These needs can include social companionship or entertainment, which try to elicit social responses from people, such as Honda humanoid, Kismet,[3] and Sony Aibo.[4] The continuum continues toward the development of systems that draw upon social attitudes to address specific needs of people, such as care-giving in healthcare;[5] autonomous systems such as in response to AAAI Robotics Challenge,[6] and "human-like" personal assistance systems such as ISAC and Cog.[7,8] This area utilizes studies in interpersonal interaction for application to interactions between people and systems. Studies have shown that people respond to artificial systems with an unconscious similarity to similar interpersonal situations, including a tendency to anthropomorphize or attribute human qualities.[9,10]

In some critical (social or nonsocial) applications, a human user interacts with a robot via Graphical User Interfaces (GUIs) and controls the robot with joystick, mouse, or similar devices. GUIs usually contain standard components considering a large number of users. Some of these user interface components may be redundant and sometimes confusing for some of the users depending on the user's preferences, capabilities, and the context in which robots are used. In addition, the users may sometimes need to control robots without any physical effort. For example, it may be hard for a disabled person to control a robot; a manual pointing device and vocal interaction might be more convenient.

Spatial reasoning ability might be important in mobile robot teleoperation, especially if the robot is at a distant location from its operator.[11,12] GUIs sometimes may create heavy information load depending on the nature of the task and the user's skills such as his/her spatial reasoning ability. For example, sonar range information may be extremely useful for people with high spatial-reasoning ability to

navigate a mobile robot while it may be not beneficial for low spatial-reasoning ability people. People with low spatial-reasoning abilities may make use of a detailed status report while people with high spatial-reasoning abilities may not.[13]

Intelligent User Interface (IUI) design has been studied in different areas including educational systems, intelligent support systems, and information filtering.[14–16] IUIs should be able to employ intelligent techniques. User adaptivity and user modeling are two of such important techniques.[17] In this research, we make use of user adaptivity and user modeling techniques. We define an adaptive user interface for robotics systems as: *"A knowledge-based interface that changes its contents to accommodate individual differences, preferences, and to reflect the mission robots are used for."* An IUI adapts itself and makes communication decisions dynamically at run-time.[18,19] An IUI differs from direct manipulation interfaces, where the former takes decision on behalf of the user and latter represents the case where the graphical objects are presented to the user for direct manipulation.[20] The architecture of IUIs includes learning the user model and inferring from the model to make decisions. The user models are extracted from the knowledge bases. Knowledge bases are structures that represent the intelligence of these interfaces. In the work of Cook and Kay the user model is displayed as a graph.[21] Each node is marked as known/not known or believed/not believed and thereby the node probabilities are inferred from the model.

Run-time adaptation of information is the key in designing IUIs. An algorithm for run-time adaptation is proposed by Gorniak and Poole.[22] This algorithm predicts future action by observing the length of the sequences of actions, the actions themselves, and the frequency of actions for predicting the future user behavior. The Incremental Probabilistic Action Modeling (IPAM)[23] is another algorithm that predicts the next element in a sequence based on detection of action patterns. Gajos *et al.* implemented three GUIs and evaluated them by comparing to a nonadaptive base. They employed recency-based and frequency-based algorithms.[24]

Speech is the main communication means for human beings. When people lack a common language, cooperation is often greatly reduced. Stating this fact, we believe users would interact with voice-controllable GUIs more efficiently than with the traditional ones. In addition, a user may need to control robot(s) without any physical effort. For example, a soldier may not be in a suitable position to command soldier robots manually or a disabled person might find vocal communication more convenient. Oviatt *et al.* discusses adaptive conversational (social) interfaces and compares them to command interfaces.[25]

This paper describes the design, implementation, and testing of a voice-controllable adaptive user interface for a mobile robot in navigational tasks. The interface offers different GUI components for a group of users depending on their capabilities, preferences, and the part of the task that they are interested in. The interface learns the users' capabilities and preferences in time as they interact more with the robot.

This paper is organized as follows: Section 2 describes the development platform and the GUI used for HRI. The voice-controllable IUI design and implementation is explained in Section 3. The experimental procedure to assess the effectiveness of the IUI is given in Section 4. The experimental results are presented in Section 5. Some conclusions are given and the future work is motivated in Section 6.

Fig. 1. Pioneer with a laptop computer attached.

2. System Architecture

The Pioneer 3-AT produced by ActivMedia is shown in Fig. 1. It has 16 sonar sensor range finders, a laser range finder, a pan-tilt-zoom camera, bumpers, and optical encoders. Fuzzy logic-based behaviors have been developed and converted into Microsoft's Component Object Model (COM) components so that they can be easily integrated. Some of the behaviors are *emulating*, *tracking, following wall, following center, move to point,* and *shadowing*. Figure 2 displays a simple GUI that is developed for voice-controllable or nonvoice-controlled (manual control) interaction with the robot. It provides drive commands, camera display with pan-tilt control, sonar and laser range finding visual displays, robot behavior controls, and status reports. Figure 3 illustrates the system architecture. The user can interact with the interface by speaking. The speech is converted to commands that are understood by the robot. The interactions of the user with the interface are recorded in a database. When the database collects sufficient metrics, the learning algorithm (described in the next section) forms a tree-structured user model. The interactions of the user with the interface are queried against the model and the system predicts the future actions based on the model.

3. Intelligent User Interfaces

An interface is made intelligent by inferring from the user model. One of the ways of developing the model is to collect the metrics of users' interaction with the interface. The metrics are saved into a database and can be retrieved when the application starts. After collecting the metrics, a user model is developed using the learning algorithms of Bayesian networks from data. Heckermen *et al.* combined the prior knowledge of user with the incoming (statistical) data to generate one or more Bayesian networks.[26] Cheng *et al.* employed an information theoretic dependency analysis for learning Bayesian network structure.[27] A message-passing algorithm for inference in Bayesian networks was developed by Pearl.[28,29]

https://doi.org/10.1017/S0263574707003414 Published online by Cambridge University Press

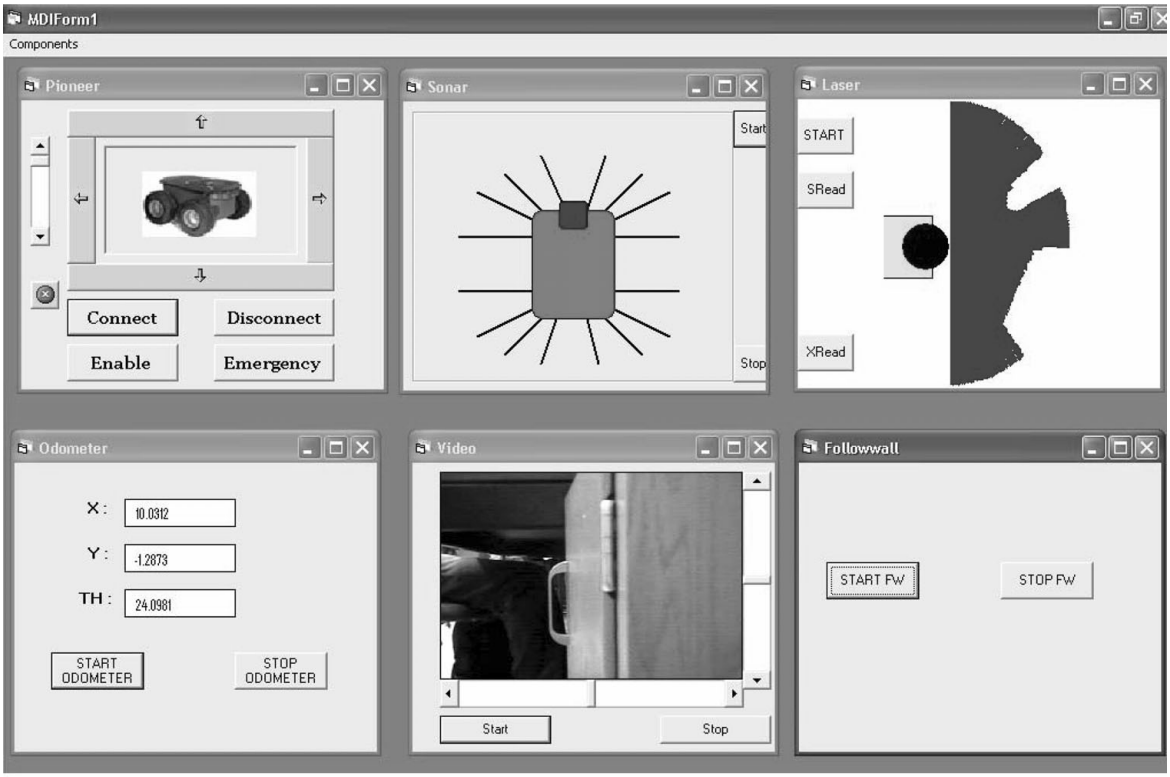

Fig. 2. Graphical user interface.

3.1. Learning algorithm

In this research, the polytree construction algorithm proposed by Rebane and Pearl is used.[30] The overall learning system is illustrated in Fig. 4. This algorithm creates polytrees from statistical data. Polytrees are singly connected networks where a child can have more than one parent, but there are no loops in the network. The dependency between the nodes with the highest score forms the link first. In this way the links between the nodes are formed in the descending order. Two kinds of data structures are considered to represent the node information. The first holds the node name and values of different calculations performed in the learning algorithm. The second holds the metrics of user's interactions with the interface. Each interaction such as opening and closing different windows of the interface is considered a metric and incremented accordingly. The different GUI components

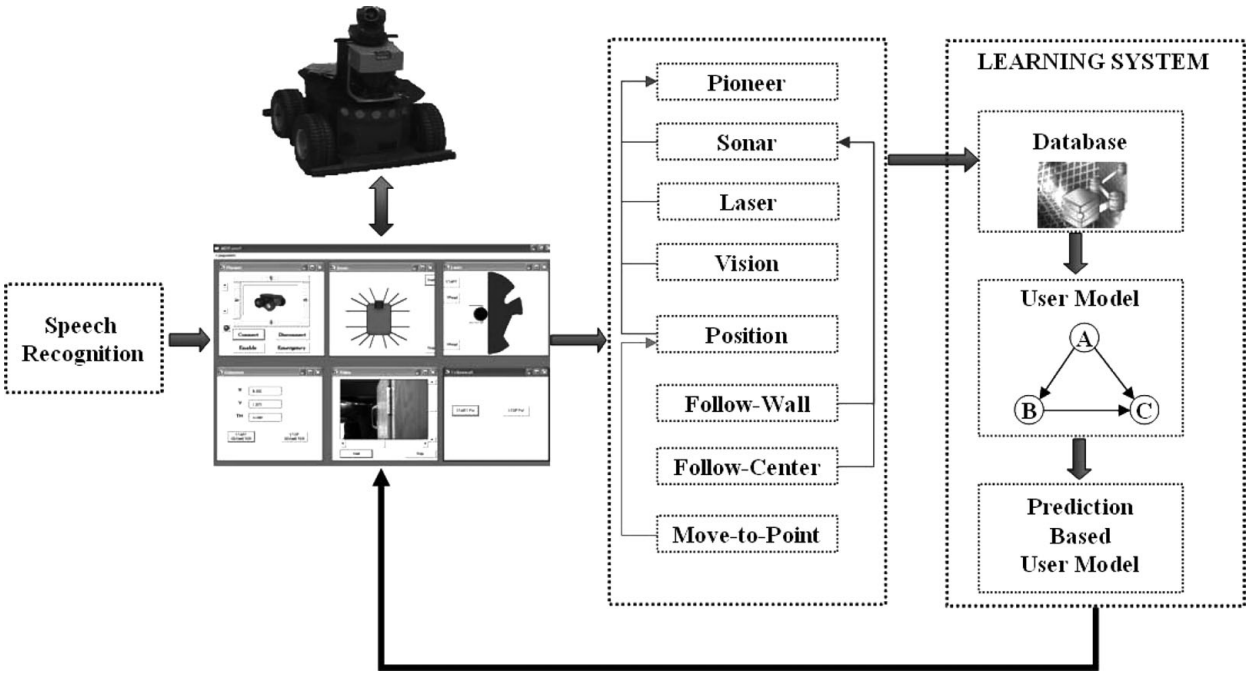

Fig. 3. System architecture.

https://doi.org/10.1017/S0263574707003414 Published online by Cambridge University Press

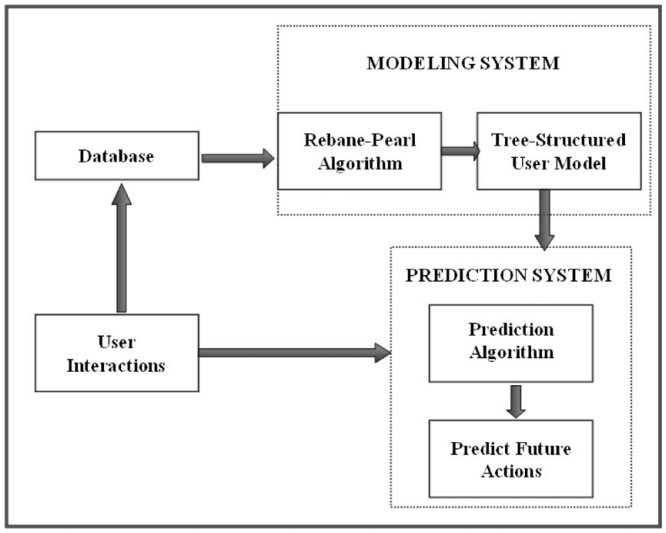

Fig. 4. Learning system.

of the interface are referred as nodes. A total of 10 nodes (pioneer, sonar, laser, odometer, video, follow-wall, follow-center, mapping, trace, and tracking) are declared. Each node has two states, open (O) and close (C). The states correspond to whether the elements on the user interface are open or closed. For example, if sonar and laser are open together, then the sonar/laser metric "OO" would be incremented. If sonar is open and laser is closed then the metric "OC" is incremented, where "OC" stands open, close. If sonar is closed and laser is open then the metric "CO" is incremented, where "CO" stands for close, open. If both sonar and laser are closed then the metric "CC" is incremented, where "CC" stands for close, close.

Once the Bayesian network is formed, the interface infers from the network when the variables are instantiated. The process of instantiating the network is by entering evidence from the interface. Evidence is the truth about a particular interface component. Supposing that the user opens the sonar form and starts sonar, then it is a true event and it is sent as evidence to the network, with a probability of 1 and sonar is said to be instantiated. If the user stops sonar, then it is sent to the network as evidence with a probability of 0. Once a node is instantiated it has to inform some of its parents and children, and the entire network is updated with the information. After the information is updated, the posterior probability of all the nodes in the network is recalculated and the node with the highest posterior probability is predicted as the users' next action.

3.2. Speech recognition

Speech is a natural communication mechanism for humans and it is hypothesized that integration of voice-control into HRI interfaces would make humans more comfortable during the interaction. In fact, voice-based interaction can be a particularly important accessibility option for some people with limited manual dexterity. In this research, Microsoft Speech Software Development Kit (SDK) 5.1 was used to develop a voice-control mechanism for HRI. The Microsoft Speech Application Programming Interface (API) provides some useful functions that are easy to integrate into the component-based software architecture as described before. All of the user interface components are COM components that are suitable to be combined with other COM components provided by Microsoft Speech SDK.

The first step for integrating speech processing is to create a speech library in eXtensible Markup Language (XML) format. The following is a sample XML file written for speech recognition:

```
<RULE NAME="Form"1ID="RID_Form1"TOPLEVEL="ACTIVE">
  <P> Pioneer </P>
      <L>
    <P> Top Right </P>
    <P> Start </P>
    <P> Stop </P>
    <P> Top Left </P>
    </L>
  </RULE>
```

The name of the form to be recognized is placed in the <P> and </P> tags. Then the control to be performed on the form is placed within list tags <L>. The list tags are used when the commands are many. Before recognizing any of the commands within the list tags, the command within the first <P> tags must be recognized. Thus the user has to say "Pioneer Start" to start the robot and "Pioneer Top Right" to place the robot on the top right corner. Using the above procedure, a single form can be placed at six different positions on the interface. The positions are the top right, left, and middle, as well as the bottom right, left, and middle. The text recognized from the speech recognition is passed to this procedure. The form is placed in the position specified by the user. Other options include closing, opening, maximizing, and minimizing the forms. The procedure checks for the word "Top" or "Bottom" first. Then it searches for the words like "Right," "Left," or "Middle". If a match is found, then the form is placed in that particular position. Thus the form can be placed in six different positions. While placing the form in one of the positions, the width and height of the form is specified so that the form does not occupy most of the space on the interface.

4. Experimental Procedure

4.1. Design

There are two conditions in the experiment: the manual control and the speech control of the interface. Spatial abilities of all the participants were assessed before the experiments and used as a continuous variable. Independent measures are the spatial ability and the number of actions to complete the given task. Dependent measures are the number of steps to complete the route and time to complete the task. There were 10 participants in each condition, randomly assigned to one of the two groups.

4.2. Participants

A total of 20 volunteers participated in the study. All volunteers are graduate/undergraduate students from the College of Engineering, Technology, and Computer Science at Tennessee State University. All participants have the same

https://doi.org/10.1017/S0263574707003414 Published online by Cambridge University Press

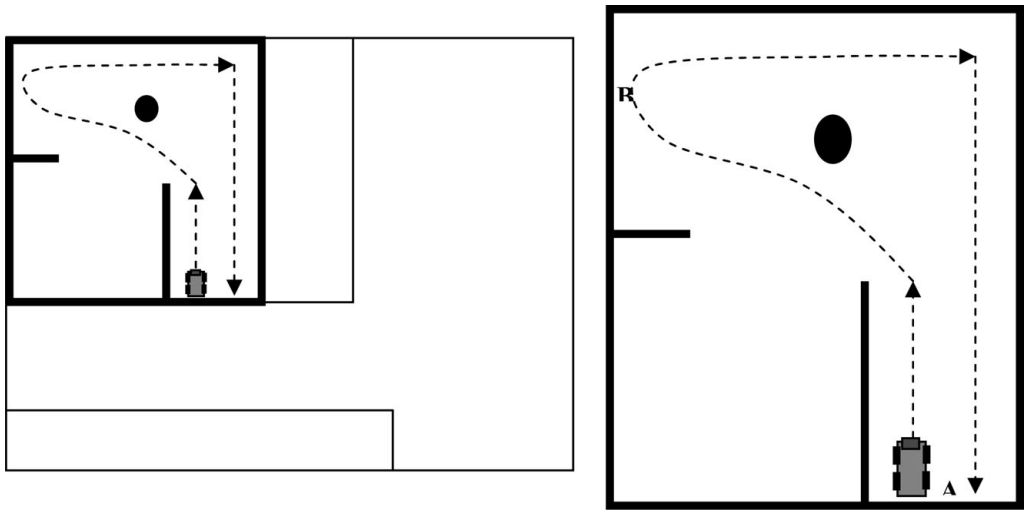

Fig. 5. Task given to the users for training.

prior knowledge of the environment in which they have to navigate the robot. To ensure this, they were asked to walk from an initial point to multiple target points until they can do it under a certain amount of time. They were then asked to draw a simple map of the environment.

4.3. Procedure

The experiment procedure involves two phases. In the first phase, the subjects were trained in using the robot. They gained experience in navigating the robot from the manual control interface. The interface has different robotic components, such as drive functions, sonar, laser, odometer, and camera. The subjects were given an environment where they have to navigate the robot using the various robotic components and behaviors. The subjects did not see the environment directly as they were only shown a map of the environment of the task. Figure 5 represents the navigation task that was used to train the subjects. 10 subjects for each control were randomly selected among 20 subjects and were provided with experiences walking around in the space in which they would later be asked to navigate the robot. Following their familiarization visit they were asked to draw a map of the space to make sure that they have equal amount of prior knowledge.

In the second phase, the subjects were given a navigation task as shown in Fig. 6. They did not see the environment directly and they were only shown a map of the environment in which they had to navigate the robot. The data collected from the task includes the number of steps (movement actions by the robot), number of actions (e.g., number

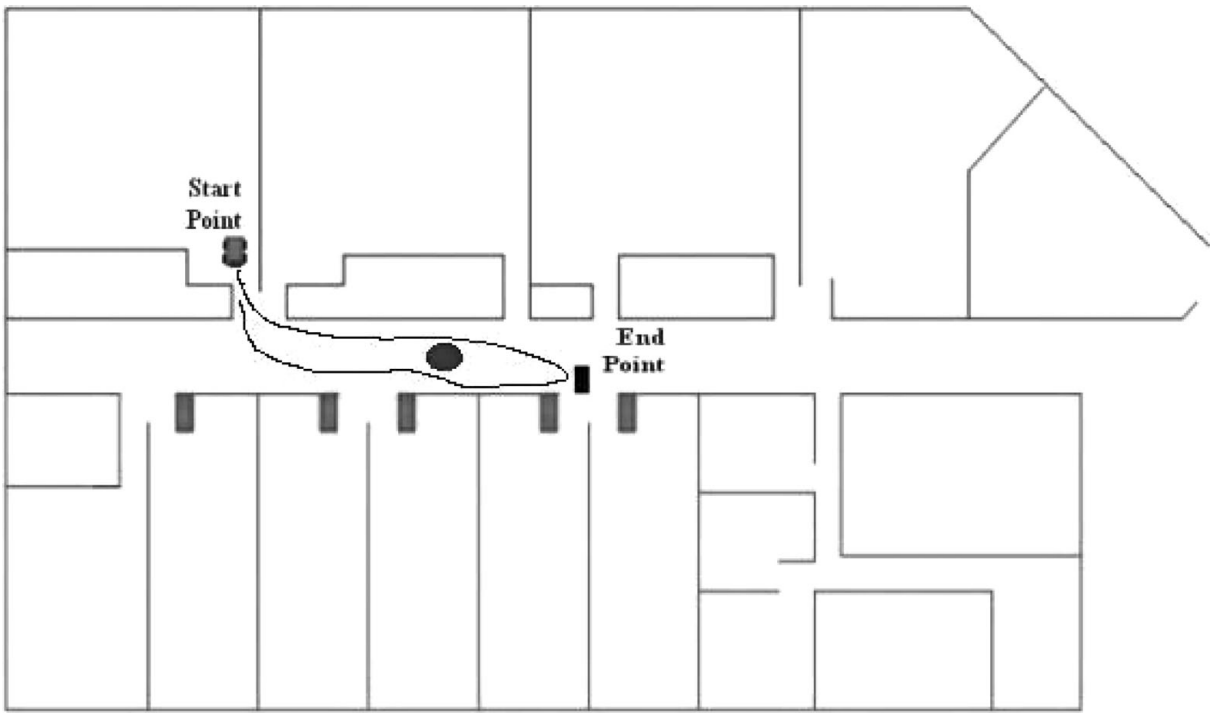

Fig. 6. Navigation task.

https://doi.org/10.1017/S0263574707003414 Published online by Cambridge University Press

of mouse clicks), and time taken for the completion of task. 10 subjects were asked to use the manual control of the robot from the interface. The other 10 subjects were asked to control the robot using speech commands. In this experiment, the subjects that used the speech control had to undergo training with the speech recognition software. The GUI that was provided to the subjects is shown in Fig. 2.

Before the subjects started navigating the robot, the Vandenberg Mental Rotation (MRT) test (revised version) was given to them to measure their spatial reasoning abilities. The test has 20 multiple choice questions and each question consists of a target figure and four possible figures two of which represent the given figure when rotated. The subjects are required to find the figures that match the target figure. Unless two of the answers are correct, no score is given for that multiple choice items. The test score can range from 0 to 100. The test involves mentally visualizing and rotating images, a skill important for spatial reasoning. After the tests, the subjects were asked to complete a brief survey of 10 questions. The survey has questions about the interface, speech, and manual control.

5. Results

5.1. Robot navigation

For each participant, we computed the mean number of steps to the target location and the mean number of seconds per navigational task over two trials. This data was modeled using multiple regression techniques, with which we tested various models to examine the combination of variables that produced the best performance predictions.

Based on the data, the best fitting model for the speech control and manual control of the robot were determined. The best fitting model for the speech control is:

$$\text{Steps} = 180.11 + 0.99 \times \text{Number of Actions} - 1.54 \times \text{Reasoning}$$

The best fitting model for the manual control is:

$$\text{Steps} = 133.41 + 5.55 \times \text{Number of Actions} - 1.76 \times \text{Reasoning}$$

It is observed that with the decrease in the MRT test score (between 0 and 100), i.e., the spatial reasoning, the number of steps required to complete the given task increases. So, the number of steps and the reasoning ability are inversely proportional to each other. The effect of the number of actions in the manual control is higher than that of the speech control.

The best fitting model for the time in the speech control is given by:

$$\text{Time} = 888.78 + 22.76 \times \text{Number of Actions} - 4.04 \times \text{Reasoning}$$

The best fitting model for the time in the manual control is given by:

$$\text{Time} = 1189.03 + 35.53 \times \text{Number of Actions} - 12.49 \times \text{Reasoning}$$

It is observed that both the spatial reasoning ability and number of actions play a more important role in the manual control in terms of time spent.

5.2. Survey result

3 out of 4 subjects with spatial reasoning ability greater than and equal to 80 (out of 100) preferred speech control of the robot over the manual control. In other words, 75% of students with spatial reasoning ability greater than 80 preferred speech over manual control of the robot. 90% of subjects were not confused to see the sonar and laser together. The sonar and laser provide similar range of information that can be used to become aware of the surrounding objects. The subjects were able to use the sensory data from those sensors without getting confused. The laser has a sweep of 180 lines and sonar has a discrete set of 16 lines representing the 16 sonar around the robot, which are placed 8 at the front and 8 at the back. 100% of the students feel that the interface is not confusing to use. 40% of students changed their views on speech recognition technology from "good" to "excellent." Most of the users felt that sonar, laser, and video are useful components of the interface.

6. Conclusions

This research developed voice-controllable intelligent interfaces and performed a user study to analyze and compare two different control mechanisms—the manual and speech controls—that can be used in a typical mobile robot navigation task. The effects of spatial reasoning ability and the number of actions were investigated. It has been found that the spatial reasoning ability is an important factor in both types of control. The number of actions has a higher impact on the manual control compared to the speech control. In addition, 75% of the high spatial reasoning ability subjects preferred the use of speech control interface.

References

1. C. Bartneck and J. Forlizzi, "A Design-Centered Framework for Social Human-Robot Interaction," *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication* (Sept. 2004) pp. 591–594.
2. C. Breazeal, *Designing Sociable Robots* (MIT Press, Cambridge, MA, (2002).
3. C. Breazeal, *Sociable machines: expressive social exchange between humans and robots*, Dissertation (Department of Electrical Engineering and Computer Science, MIT, 2000).
4. M. Fujita, "AIBO: towards the era of digital creatures," *Int. J. Robot. Res.* **20**(10), 781–794 (2001).
5. M. Pollack *et al.*, "Pearl: A Mobile Robotic Assistant for the Elderly," *Proceedings of AAAI Workshop on Automation as Eldercare* (2002) pp. 85–92.
6. R. Simmons *et al.*, "GRACE and GEORGE: Autonomous Robots for the AAAI Robot Challenge," *Proceedings of AAAI 2004 Mobile Robot Competition Workshop* (2004) pp. 15–20.

https://doi.org/10.1017/S0263574707003414 Published online by Cambridge University Press

7. K. Kawamura, T. Rogers and X. Ao, "Development of a Cognitive Model of Humans in a Multi-Agent Framework for Human–Robot Interaction," *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent System*, Bologna, Italy (2002) pp. 1379–1386.
8. B. Scassellati, *Theory of mind for a humanoid robot*, Ph.D. dissertation (Department of Electrical Engineering and Computer Science, MIT, 2001).
9. S. Kiesler and J. Goetz, "Mental Models and Cooperation With Robotic Assistants" **In:** *CHI 2002 Extended Abstracts* (ACM Press, Minneapolis, MN, 2002) pp. 576–577.
10. B. Reeves and C. Nass, *The Media Equation* (Cambridge University Press, Cambridge, UK, 1996).
11. A. Sekmen, M. Wilkes, S. Goldman and S. Sabatto, "Exploring importance of location and prior knowledge in mobile robot control," *Int. J. Human Comput. Stud.* **58**(1), 5–20 (2003).
12. D. J. Bruemmer, D. A. Few and C. W. Nielson, "Spatial Reasoning for Human–Robot teams," **In**: *Emerging Spatial Information Systems and Applications* (Brian Hilton, ed.) (Idea Group Inc., 2006) pp. 350–372.
13. A. Sekmen, *Human–robot interaction methodology*, Ph.D. dissertation (Vanderbilt University, 2000).
14. A. Granic and V. Glavinic, "Automatic Adaptation of User Interfaces for Computerized Educational Systems," *Proceedings of 10th IEEE International Conference on Electronics, Circuits and Systems*, NJ, USA (2003), pp. 1232–1235.
15. D. Benyon and D. Murray, "Adaptive systems: from intelligent tutoring to autonomous systems," *Knowl. Based Syst.* **6**(4), 197–219 (1993).
16. M. H. Chignell and P. A. Hancock, *Intelligent Interfaces, Handbook of Human-Computer Interaction* (Elsevier, Amsterdam, The Netherlands, 1988).
17. K. Hook, "Steps to take before IUI becomes real," *Journal of Interacting with Computers* **12**(4), (2000), pp. 409–426.
18. P. Szekely, "Structuring Programs to Support Intelligent Interfaces," **In**: *Intelligent User Interfaces* (J. Sullivan and S. Tyler, eds.) (ACM Press, 1991) pp. 445–464.
19. C. Karagiannidis, A. Koumpis and C. Stephanidis, "Decision Making in Intelligent User Interfaces," *Proceedings of the ACM International Conference on Intelligent User Interfaces* (1997) pp. 195–202.
20. E. L. Hutchins, J. D. Hollan and D. Norman, "Direct manipulation interfaces," *Human-Comput. Interact.* **1**, 311–338 (1985).
21. R. Cook and J. Kay, "The Justified User Model: A Viewable, Explained User Model," *Proceedings of the 4th International Conference on User Modeling*, Hyannis, Massachusetts (1994) pp. 145–150.
22. P. Gorniak and D. Poole, "Predicting Future User Actions by Observing Unmodified Applications," *Proceedings of the 17th National Conference on Artificial Intelligence* (Aug. 2000) pp. 217–222.
23. B. D. Davison and H. Hirsh, "Predicting sequences of user actions," *Predicting the Future: AI Approaches to Time Series Problems*, Technical Report (1998) pp. 5–12.
24. K. Z. Gajos, M. Czerwinski, D. S. Tanb and D. S. Weld, "Exploring the Design Space for Adaptive Graphical User Interfaces," *Proceedings of the Working Conference on Advanced Visual Interfaces* (2006) pp. 201–208.
25. S. Oviatt, C. Darves and R. Coulston, "Toward adaptive conversational interfaces: modeling speech convergence with animated personas," *ACM Trans. Comput.–Human Interact.*, **3**(11), 300–328 (2004).
26. D. Heckermen, D. Geiger and D. Chickering, "Learning Bayesian networks: the combination of knowledge and statistical data," Technical Report, Microsoft Research (1994).
27. J. Cheng, D. Bell and W. Liu, "Learning Bayesian Networks From Data: An Efficient Approach Based on Information Theory," *Proceedings of the 6th ACM International Conference on Information and Knowledge Management* (1997) pp. 325–331.
28. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, San Mateo, CA, 1988).
29. J. Pearl and T. M. Verda, "A Theory of Inferred Causation," *Proceedings of the 2nd Conference on Principles of Knowledge Representation and Reasoning* (1991) pp. 441–452.
30. T. Rebane and J. Pearl, "The Recovery of Causal Poly-Trees From Statistical Data," **In**: *Uncertainty in Artificial Intelligence* (L.N. Kanal, T.S. Levitt and J.F. Lemmer, eds.) (Amsterdam, The Netherlands, 1989).