A Robotic Agent in a Virtual Environment that Performs Situated Incremental Understanding of Navigational Utterances

Takashi Yamauchi
Seikei University
3-3-1 Kichijoji-Kitamachi
Musashino, Tokyo, Japan
dm126222@cc.seikei.ac.jp

Mikio Nakano
Honda Research Institute
Japan Co., Ltd.
8-1 Honcho, Wako
Wako, Saitama, Japan
nakano@jp.honda-ri.com

Kotaro Funakoshi
Honda Research Institute
Japan Co., Ltd.
8-1 Honcho, Wako
Wako, Saitama, Japan
funakoshi@jp.honda-ri.com

Abstract

We demonstrate a robotic agent in a 3D virtual environment that understands human navigational instructions. Such an agent needs to select actions based on not only instructions but also situations. It is also expected to immediately react to the instructions. Our agent incrementally understands spoken instructions and immediately controls a mobile robot based on the incremental understanding results and situation information such as the locations of obstacles and moving history. It can be used as an experimental system for collecting human-robot interactions in dynamically changing situations.

1 Introduction

Movable robots are ones that can execute tasks by moving around. If such robots can understand spoken language navigational instructions, they will become more useful and will be widely used. However, spoken language instructions are sometimes ambiguous in that their meanings differ depending on the situations such as robot and obstacle locations, so it is not always easy to make them understand spoken language instructions. Moreover, when they receive instructions while they are moving and they understand instructions only after they finish, accurate understanding is not easy since the situation may change during the instruction utterances.

Although there have been several pieces of work on robots that receive linguistic navigational instructions (Marge and Rudnicky, 2010; Tellex et al., 2011), they try to understand instructions before moving and they do not deal with instructions when situations dynamically change.

We will demonstrate a 3D virtual robotic system that understands spoken language navigational instructions in a situation-dependent way. It incrementally understands instructions so that it can understand them based on the situation at that point in time when the instructions are made.

2 A Mobile Robot in a 3D Virtual Environment

We use a robotic system that works in a virtual environment built on top of SIROS (Raux, 2010), which was originally developed for collecting dialogues between two participants who are engaging in an online video game. As an example, a convenience store environment was developed and a corpus of interaction was collected (Raux and Nakano, 2010). One of the participants, the operator, controls a (simulated) humanoid robot whose role is to answer all customer requests. The other participant plays the role of a remote manager who sees the whole store but can only interact with the operator through speech. The operator has the robot view (whose field of view and depth are limited to simulate a robot's vision) and the manager has a birds-eye view of the store (Figure 1). Customers randomly visit the store and make requests at various locations. The manager guides the operator towards customers needing attention. The operator then answers the customer's requests and gets points for each satisfied request.

Using the virtual environment described above, we have developed a system that operates the robot according to the human manager's instructions. Currently we deal with only navigational instructions for moving the robot to a customer.

Figure 2 depicts the architecture for our system. We use Sphinx-4 (Lamere et al., 2003) for speech recognition. Its acoustic model is trained on the Wall Street Journal Corpus and its trigram language model was trained on 1,616 sentences in the human-human dialogue corpus described above. Its vocabulary size is 275 words. We use Festival (Black et al., 2001) for speech synthesis.

Proceedings of the SIGDIAL 2013 Conference, pages 369–371,
Metz, France, 22-24 August 2013.

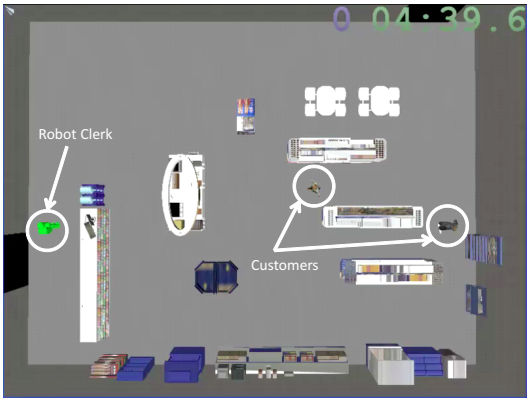

Figure 1: The manager's view of the convenience store.

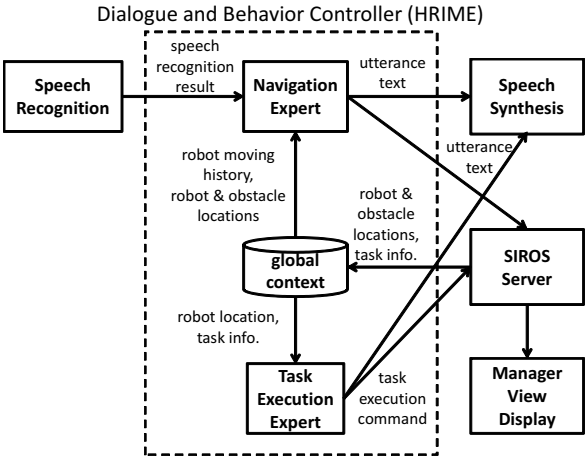

Figure 2: System architecture.

We use HRIME (HRI Intelligence Platform based on Multiple Experts) (Nakano et al., 2008) for dialogue and behavior control. In an HRIME application, experts, which are modules dedicated to specific tasks, are activated at appropriate times and perform tasks. The navigation expert is activated when the system receives a navigational instruction. There are seven semantic categories of instructions; they are *turn-right*, *turn-left*, *go-forward*, *go-back*, *repeat-the-previous-action*, *do-the-opposite-action-of-the-previous-one*, and *stop*. Utterances that do not fall into any of these are ignored. We assume that there are rules that match linguistic patterns and those semantic categories. For example, "right" corresponds to *turn-right*, and "more" corresponds to *repeat-the-previous-action*. The navigation expert sends the SIROS server navigation commands based on the recognized semantic categories. Those commands move the robot in the same way as a human operator operates the robot using the keyboard, and the results are shown on the display the manager is watching. When the robot starts moving and it cannot move because of an obstacle, it reports it to the manager by sending its utterance to the speech synthesizer.

When the robot has approached a customer who is requesting help, the task is automatically performed by the task execution expert.

The global context in the dialogue and behavior controller stores information on the environment which is obtained from the SIROS server, and it can be used by the experts. As in the same way in the human-human interaction, it holds information only on customers and obstacles close to the robot so that restricted robot vision can be simulated.

3 Situated Incremental Understanding

Sometimes manager utterances last without pauses like "right, right, more right, stop", and the situation changes during the utterances because the robot and the customers can move. So our system employs incremental speech recognition and moves the robot if a navigational instruction pattern is found in the incremental output. To obtain incremental speech recognition outputs, we employed InproTK (Baumann et al., 2010), which is an extension to Sphinx-4. It enables the system to receive tentative results every 10ms, which is a hypothesis for the interval from the beginning of speech to the point in time.

However, since incremental outputs are sometimes unstable and the instructions are ambiguous in that the amount of movement is not specified, not only incremental speech recognition outputs but also obstacle locations and moving history is used to determine the navigation commands.

In our system, the robot navigation expert receives incremental recognition results and if it finds a navigational instruction pattern, it consults the situation information in the global context, and issues a navigation command based on several situation-dependent understanding rules that are manually written. Below are examples.

- If there is an obstacle in the direction that the recognized instruction indicates, ignore the recognized instruction. For example, when "go forward" is recognized but there is an obstacle ahead, it is guessed that the recognition result was an error.

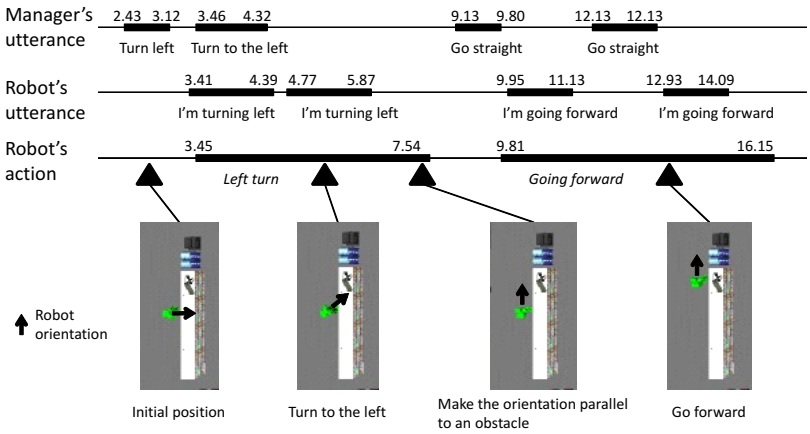

Figure 3: Interaction example.

- When rotating, adjust the degree of rotation so that the resulting orientation becomes parallel to obstacles such as a display shelf. This enables the robot to smoothly go down the aisles.

Figure 3 shows an example interaction. In the demonstration, we will show how the robot moves according to the spoken instructions by a human looking at the manager display. We will compare our system with its non-incremental version and a version that does not use situation-dependent understanding rules to show how incremental situated understanding is effective.

4 Future Work

We are using this system for collecting a corpus of human-robot interaction in dynamically changing situations so that we can analyze how humans make utterances in such situations. Future work includes to make the system understand more complicated utterances such as "turn a little bit to the left". We are also planning to work on automatically learning the situation-dependent action selection rules from such a corpus (Vogel and Jurafsky, 2010) to navigate the robot more smoothly.

Acknowledgments

We thank Antoine Raux and Shun Sato for their contribution to building the previous versions of this system. Thanks also go to Timo Baumann Okko Buß, and David Schlangen for making their InproTK available.

References

Timo Baumann, Okko Buß, and David Schlangen. 2010. InproTK in Action: Open-Source Software for Building German-Speaking Incremental Spoken Dialogue Systems. In *Proc. of ESSV*.

Alan Black, Paul Taylor, Richard Caley, Rob Clark, Korin Richmond, Simon King, Volker Strom, and Heiga Zen. 2001. The Festival Speech Synthesis System, Version 1.4.2. *Unpublished document available via http://www. cstr. ed. ac. uk/projects/festival. html*.

Paul Lamere, Philip Kwok, William Walker, Evandro Gouvea, Rita Singh, Bhiksha Raj, and Peter Wolf. 2003. Design of the CMU Sphinx-4 decoder. In *Proc. of Eurospeech-2003*.

Matthew Marge and Alexander I. Rudnicky. 2010. Comparing spoken language route instructions for robots across environment representations. In *Proc. of SIGDIAL-10*.

Mikio Nakano, Kotaro Funakoshi, Yuji Hasegawa, and Hiroshi Tsujino. 2008. A framework for building conversational agents based on a multi-expert model. In *Proc. of SIGDIAL-08*, pages 88–91.

Antoine Raux and Mikio Nakano. 2010. The dynamics of action corrections in situated interaction. In *Proc. of SIGDIAL-10*, pages 165–174.

Antoine Raux. 2010. SIROS: A framework for human-robot interaction research in virtual worlds. In *Proc. of the AAAI 2010 Fall Symposium on Dialog with Robots*.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proc. of AAAI-2011*.

Adam Vogel and Dan Jurafsky. 2010. Learning to follow navigational directions. In *Proc. of ACL-2010*, pages 806–814.