Designing a Multi-modal Language for Directing Multipurpose Home Robots

Tetsushi Oka
Fukuoka Institute of Technology
oka@fit.ac.jp.

Masao Yokota
Fukuoka Institute of Technology
yokota@fit.ac.jp

Abstract

This paper discusses an approach to realising multipurpose home robots a wide spectrum of people can direct by speech, physical contact and gesture. A good spoken language interface allows ordinary people to direct robots without training in advance. However, many problems of Natural Language Understanding must be addressed if a wide range of utterances are to be understood immediately. Therefore, we design a practical multi-modal direction language combining a simple spoken language with nonverbal information. The spoken language has a simplified grammar and limited words, so that the robot should understand commands without complex computation or a large knowledge base. Nonverbal information makes commands more specific and eliminates ambiguity.

1 Introduction

Since the end of last century, more and more robots are coming into homes and offices to help ordinary people. As the birthrates in advanced countries are dropping, a robot that helps and cares elder or disabled people at home will be soon in demand in many societies. Such a robot is expected to understand what the user wants it to do as soon as a command is given. This raises an important issue of human-robot communication. For people who need helps, even computer GUIs, or remote controls of TVs are not ideal interfaces. Some people may give up using the robot before learning which commands all the buttons, sliders and levers are linked with. In addition, if the robots are to execute many kinds of commands, the user must learn long sequences of operations.

A spoken language interface, on the other hand, can be a good interface as it is not necessary to learn a new language if robots understand spoken commands in our language. However, there are still many difficult problems to tackle to realize a natural language interface by speech. For example, a spoken language interface based on a linguistically motivated grammar, compositional semantics of logical forms and generic inference engines for natural language understanding has been applied to directing a mobile robot, asking it questions and giving it information by speech[1]. This system requires considerable computational power to check the consistency of semantic interpretations of user utterances in real time.

Recently, many robot systems implementing automatic speech recognition have been developed[2]. Many of them create word lattices from speech signals and interpret what the user means by means of keyword spotting, creating semantic representations of user utterances. In most systems, heuristic rules to construct semantic representations are built by the designer so that it should not take much time for semantic analysis. However, it is not clear to users what kind of an utterance the robot understands or misunderstands. Moreover, it is difficult to identify the meaning by this method alone if a variety of commands are given to the robot in many different situations. Therefore, this approach does not suit for multipurpose robots which help ordinary people at home.

Another way to reduce the computational cost of semantic analysis is to employ a simple command language. The most straightforward way is to design a command language which robots can directly execute, but the language would be difficult for humans to learn. Therefore, we design a command language based on a natural language by reducing the number and types of grammar rules, and the size of the lexicon. Obviously, a command language does not need to cover declarative and interrogative utterances. We presume that it is possible to build a practical and cost-effective spoken language interface by selecting words and phrases and constructing rules carefully. Besides spoken commands, gestures and physical contact provide a good means of intuitive direction. Nonverbal modalities of communication complement verbal communication and eliminate ambiguity. Thus, we design a multi-modal language combining a spoken language with nonverbal directions.

2 Target of the language

2.1 Multipurpose Home Robots

Home robots including pet robots like Sony AIBO (http://www.aibo.com/) and healing robots, PARO (http://www.paro.jp/) for instance, are already in the market and helping ordinary people at homes in some ways. Another example is a practical small robot available for cleaning our rooms (http://www.irobot.com). Although it is still difficult to build a robot which replaces a housekeeper or care worker, one can predict that home robots will evolve in the coming decades and be capable of doing many kinds of tasks given by us. They will be connected to home computer networks and collect useful information for us. They will move about in the home, bring something to us and move heavy objects. They will help us doing physical tasks, have a chat with us, control the air conditioner, TV and lights, and so forth. In short, we predict multipurpose home robots will find a place to help us.

In a way, a home robot can be thought of as a physical interface device with a home computer network or an intelligent house which follows the user and provides a means of interaction with the computers. At the same time, the robot is given physical tasks and helps the user. Most importantly, the robot has sensors and actuators and can change its location. This makes the way it interacts with a human utterly different from other interface devices.

2.2 Directing Mobile Robots

We believe that no matter how intelligent the robots may be, in some cases we have to direct them step by step telling them to stand up, turn back, go forward, look left, raise the right arm, grasp an object and so on, especially when the robots are physically helping them. Thus, the first step would be designing a multimodal language which enables us to make robots turn, move forward/backward, look up/down and stop at will, saying "Turn left!", touching the robot, waving to it and so on.

It looks straightforward to realize a simple spoken language for this purpose, but things get a little complicated when we want to specify parameters of actions such as distance, angle and speed. On the one hand, "Turn right slowly!" does not include detailed information about the angle and speed, but on the other hand, "Turn 43 degrees clockwise within a second!" is not what we normally say. We could say "stop there", "faster", "a little bit more" after "go forward slowly" or "turn left a bit" many times until the robot reaches a desirable position. Nonverbal signals such as gestures can allow the user to give the robot detailed information in a more natural manner. Thus, a multimodal language combining a simple spoken command language with gestures and physical contact opens up possibilities to direct robots in natural ways.

The spoken language can be defined by a set of simple grammar rules and a relatively small lexicon. Besides, one can realize a wider coverage of commands to home robots using the same grammar rules and adding words or phrases to the lexicon: "Turn on the TV!", "Go to the kitchen", "(Make the room) warmer!", "Clean up the room", "Check my mailbox!", "Show me the weather forecast!", "Lift the box" and so on. Since in most of the environments of home robots, one can assume that e.g. there is only one TV in the room the user is in, little reasoning will be necessary to understand the commands if we properly design the language.

2.3 Directing Articulated Robots

Articulated robots with arms, hands and legs can perform many kinds of physical tasks. When using such robots, we will often want to tell the robot move its hands, arms or legs: "Stand up!", "Sit down!", "Walk a little!", "Raise the right arm higher!", "Stand on the left foot!", "Wave the arms slowly!", etc. These directions specify either a primitive action using the arms and legs or movements of individual limbs, so little confusion occurs in order to generate motor commands to the actuators, although a robust balance control system is indispensable.

As a humanoid has many degrees of freedom, the diversity of its motions is much larger than that of a wheeled mobile robot. Thus, we need a wider coverage in our multi-modal language. None the less, it is possible to adapt the language for this purpose by enhancing the lexicon without adding grammar rules.

If we could easily direct humanoids' various motions in a multi-modal language, it would be possible to make them help us in physical tasks and teach them how to move their limbs.

3 Spoken Language for Intuitive Direction of Home Robots

3.1 Grammar and Lexicon

Now, we discuss in more detail what kind of a spoken language is suitable for directing home robots.

What we have principally in mind is a grammar consisting of a small number of rules without recursions. For our purpose, we can select words necessary for directing robots. We presume that this is the class we should employ for a wide coverage of commands to multipurpose home robots. The grammar restricts the number of acceptable commands, but compositional semantics makes it to possible to interpret a large number of spoken commands and convert them into robot actions. However, even a grammar in this simple class generates unnecessary utterances, so there are utterances which are grammatical but do not make sense. Only utterances which are grammatical and make sense should be converted into robot actions and the other grammatical input must be recognised as invalid at the stage in which their semantic representations are constructed. Thus, the robots need react to spoken commands in three, at least, different ways.

There are many ways to define a language of this class to cover verbal directions. Designing our language, we do not adhere to linguistic grammars and lexical categories of natural languages. For our purpose, it is more important to consider to what extent we should restrict directions to minimise the cost of computation and how easy for humans the language is to learn. Here are some examples of grammar rules:

$$
\begin{aligned}
S &\rightarrow \mathit{ACTION}\ \mathit{PARAM} \\
\mathit{ACTION} &\rightarrow \mathit{turn} \\
\mathit{PARAM} &\rightarrow \mathit{DIRECTION}\ \mathit{SPEED} \\
\mathit{DIRECTION} &\rightarrow \mathit{ANGLE}\ \mathit{left} \\
\mathit{ACTION} &\rightarrow \mathit{OPERATION}\ \mathit{OBJECT} \\
\mathit{OPERATION} &\rightarrow \mathit{turn}\ \mathit{off} \\
\mathit{OBJECT} &\rightarrow \mathit{the}\ \mathit{TV}
\end{aligned}
$$

Note that without recursions separate rules are necessary to allow the users to specify different combinations of parameters.

$$\mathit{PARAM} \rightarrow \mathit{DIRECTION}\ \mathit{DURATION}$$

3.2 Semantic Analysis and Robot Actions

Once the syntax of our language for verbal directions is defined, we need to give a meaning to each utterance. For example, if "Move!" is grammatical, we need to decide whether it is acceptable as a command or not and what actions it should be linked with depending on nonverbal input. Although generally speaking the meaning of a command can vary depending on the context, we should avoid allowing such an ambiguous command. Instead, we should be able to choose one action, if an acceptable combination of an utterance and nonverbal input is given.

If the user says "Turn right!", the robot must change its orientation though it is not clear how much and how fast without nonverbal input. Perhaps the robot should turn, say, 45 degrees clockwise at a moderate speed. To execute an action, it needs all the parameters of it. For our purpose, the language interface should send action representations to the robot control system whenever the user commands the robot. A simple example of action representations passed to the robot system should look like the following:

```
action(name(turn),dir(45.0),speed(50.0),time(now))
```

Semantic analysis is necessary to construct action representations from verbal and nonverbal input. Methods of compositional semantics can systematically construct semantic and action representaions from speech input. We believe that it is important to have a good framework for semantic analysis of multimodal directions. Mapping utterances into actions in an ad hoc manner is not desirable for building a language for diverse directions. The Mental Image Directed Semantic Theory (MIDST) gives a framework for semantic processing in multi-modal interactions between a human and an articulated robot[3], and thus will give one for our purpose.

3.3 Disambiguation

As mentioned above, ambiguity in spoken commands should be avoided as much as possible. First, we can eliminate context dependent directions such as "Go *there*!" and "Approach *it*". Secondly, we choose default values of action parameters. For example, we could link "Turn!" and an action to turn 360 degrees slowly, or "Pick *it* up!" and an action to grasp anything near the robot and hold it. Thirdly, an utterance can be interpreted in more than one way because of multi-sense words or syntactically ambiguous expressions. Although most of such utterances are ruled out if the language is simple, we must consider this problem at both of the stages of designing the grammar and lexicon and mapping utterances into actions.

4 Nonverbal Direction

Home robots will have many different sensors for perceiving their circumstances and controlling their

bodies. Those sensors, especially tactile, proximity, and vision sensors, are useful as user interface devices. In our multi-modal language, nonverbal input through sensors plays an important role for disambiguation, filling slots of action representations. Pointing gestures disambiguate directions like "Go *over there*!" and "Touch *that* wall!". Tapping a part of the robot implies the direction, speed and duration of the action: saying "Turn!" and tapping the robots left arm three times can be interpreted as "Turn 90 degrees to the left quickly!".

5 Robot Systems for Usability Studies

For usability studies, we build robot systems one can direct in our multi-modal language and conduct experiments involving people who are not familiar with computers and robots.

The robot systems consist of a multi-modal language interface and robot control system. The language interface receives sound signals and other sensory input. It comprises components for speech recognition, gesture recognition, analysis of tactile and proximity sensor readings, and syntactic and semantic analysis. It sends the robot control system action representations which describe an action directed by the user in real time .

We are currently implementing robot systems with a multi-modal command interface. Our first prototype was built on a Sony AIBO (ERS-210) using the *Master Studio SDK*. The robot recognises about 50 Japanese words, simple hand gestures moving a pink ball and tactile messages on its tactile and infrared sensors. Its action repertoire includes standard action of the SDK, standing up, walking etc., and our original actions created using *Action Composer*, a part of the SDK. Although it can only react to single-word commands, it is possible to realise an intuitive interface to direct the robot. The major limitations are due to the small sets of spoken commands and nonverbal directions.

Our next setup, Lemon (Fig.1), is based on a new Sony AIBO (ERS-7M3), the *Open-R SDK* (http://openr.aibo.com/) which enables us to build the on-board control system in C++ and *MEdit* to create robot motions to add new actions. For speech recognition, we use *Julian* (http://julius.sourceforge.jp/en/julius.html), a grammar based recognition engine and its development kit which allows us to develop context free grammars[1] for speech input and test them on our PCs. We develop a spoken language based on Japanese using Julian and study its usability on our robot system.

Another target is a humanoid robot one can direct using our multi-modal language. We are developing small humanoids for various objectives. The newest at the moment is Syokabe (Fig.1), which has 27 DOFs including effective yawing rotations at the body and legs. Running Julian on our PCs, we can use the same spoken language interface for Lemon and our humanoids.

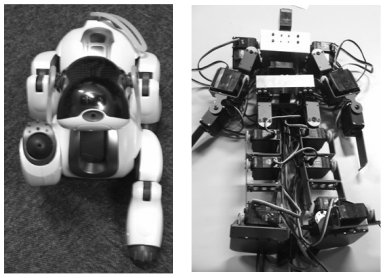

Figure 1: Lemon (left) and Syokabe

6 Summary

This paper proposed to design a practical multi-modal language to direct multipurpose home robots. We discussed basic features of the language, action representations, and command interface. The language is based on a natural language and has a smaller set of grammar rules to cover directions of home robots. We have implemented some prototypes and components of robot systems and are conducting usability studies using four legged robots and humanoids.

References

[1] J. Bos and T. Oka "A spoken language interface with a mobile robot" *Journal of Artificial Life and Robotics*, Vol. 11, 2007.

[2] R. Prasad, H. Saruwatari and K. Shikano, "Robots that can hear, understand and talk" *Advanced Robotics*, Vol. 18, pp. 533-564, 2004.

[3] M.Yokota and G.Capi, "Integrated multimedia understanding based on Mental Image Directed Semantic Theory" *Proc. 11th International Symposium on Artificial Life and Robotics*, OS6-1, 2006.

[1]The engine actually parses only regular languages.