A new human-robot interaction method using semantic symbols

Sang-Hyun Park*, Jung-Hoon Hwang*, and Dong-Soo Kwon**

* Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Deajeon, Korea
(Tel : +82-42-869-3082; E-mail: {park_sh, hwangjh}@robot.kaist.ac.kr)

**Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Deajeon, Korea
(Tel : +82-42-869-3042; E-mail: kwonds@kaist.ac.kr)

Abstract: As robots become more prevalent in human daily life, situations requiring interaction between humans and robots will occur more frequently. Therefore, human-robot interaction (HRI) is becoming increasingly important. Although robotics researchers have made many technical developments in their field, intuitive and easy ways for most common users to interact with robots are still lacking. This paper introduces a new approach to enhance human-robot interaction using a semantic symbol language and proposes a method to acquire the intentions of robot users. In the proposed approach, each semantic symbol represents knowledge about either the environment or an action that a robot can perform. Users' intentions are expressed by symbolized multimodal information. To interpret a users' command, a probabilistic approach is used, which is appropriate for interpreting a freestyle user expression or insufficient input information. Therefore, a first-order Markov model is constructed as a probabilistic model, and a questionnaire is conducted to obtain state transition probabilities for this Markov model. Finally, we evaluated our model to show how well it interprets users' commands.

Keywords: Human-robot interaction, semantic symbols, user command, probabilistic model

1. INTRODUCTION

Today, robots can be found frequently in human daily life, and this trend is continuing. There are predictions of a one robot per household revolution, analogous to the advance of the ubiquitous PC. In such a world, there would be clear advantages to robots that have some capacity to think like humans.

During human-to-human interaction, it is obvious that using various modalities is natural and easy for humans communicating with one another. For this reason, many researchers emphasize the key role of multimodal information in human interactions [1, 2]. A multimodal system involves two or more combined user input modes – such as speech, touch, manual gestures, gaze, and head and body movements – that are processed in a coordinated manner with a multimedia system output [3]. In 1980, Bolt proposed a "Put that there" system [4]. This system was the first system that processed a speech and pointing gesture in parallel. In this system, a user can make a voice command and he or she can specify what "that" and "there" refer to by pointing to a certain object or a place. Since the development of this system, robotics researchers have been trying to develop a similar system for human-robot interactions. Bischoff *et al*. designed the humanoid robot HERMES [5]. In the HERMES system, inputs are conveyed to the robot via voice, keyboard, or e-mail. The transmitted inputs are combined and translated using natural language processing technology. The system separates the character input string into a sequence of words and numbers, and then gave the parsed-words a part of speech, finally compare them with a list of prototype command sentence. Another robot that can manipulate multimodal information is , developed by Perzanowski *et al*. [6]. Operating this robot, users can combine speech, gesture, and PDA inputs in various ways. Although robots like HERMES and Coyote use multimodal information, present HRI research focuses mainly on the processing of natural language to facilitate communication with humans. A complementary relationship involving multimodal information is a main research topic in this area.

In previous studies, little attention has been paid to situations in which a freestyle input sentence or insufficient input information is given to the robot. Actually, many situations arise in which insufficient or inadequate input information (as interpreted by the robot) is given. For example, if we want a robot to fetch a cup, we can give commands such as these: "Cup," "Fetch a cup," or even "Cup Fetch" (In Korean, this would be correct grammar.). Although the "Cup" command should be supplemented to make the meaning of the command clear, a person might correctly interpret the command according to the context and go to fetch a cup. If a robot always requests supplementary information whenever it receives insufficient input or when a user states a command in an ungrammatical way, the human user will soon feel annoyed and uncomfortable.

An ideal robot would have the ability to understand a great variety of user expressions. Although perfect understanding of all kinds of user input is difficult to achieve, a human user wants a robot to be able to understand his or her exact intention. A robot also should understand a user command whether it is grammatically correct or not. To enable a robot to have such understanding, a semantic symbol-based human robot interaction method is proposed. By representing multimodal information as semantic symbols, the user's intentions can be converted into a form that can be understood by the robot. These semantic symbols are given to the robot in word level meaning. In section 2, we will discuss the meaning of the semantic symbols: the symbols represent the objects and places in the environment and the actions the robot is able to carry out. At the end of section 2, tasks which a robot can perform in the assumed environment will be referred. In section 3, we suggest a probabilistic model. The raw users' inputs can be interpreted as specified tasks using these models. In addition, the overall structure will be mentioned. In section 4, we will discuss the results of our evaluation.

2. SEMANTIC SYMBOLS AND TASKS

2.1 Semantic symbols

In the proposed semantic symbol-based human robot interaction method, a semantic symbol is a basic element of a user command. It is assumed that word-level semantic symbols are obtained from visual information or from voice commands. These symbols represent physical objects, actual places, and robot actions. Figure 1 shows this relation. The

semantic symbols represent some meaning, rather than vocal or visual features, that is, they are abstracted to have meaning. This paper is written under the supposition that an abstraction that converts the features to symbols has already been implemented. When the semantic symbols are put successively to the robot, these semantic symbols make a semantic symbol sentence. This paper focuses on a method of abstracting the user's intention (expressed by semantic symbol sentence) to command tasks that are known by the robot.

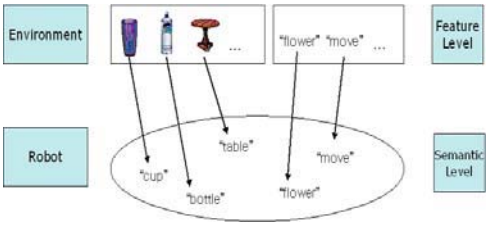

Fig. 1 Multimodal information and semantic symbol

2.2 Robot actions and environment

We assume that robot actions are as follows:

- Moving from one place to another place
- Grasping an object and releasing it
- Seeing an object or a place.
- Pressing a switch

The first our test bed for HRI is the mobile robot having one manipulator and it provides the various services to a human in his or her home. Semantic symbols are cautiously chosen for this limited domain, and these symbols should be assigned to each action (Move, Grasp, See, Release, Press, and Return), according to the meaning of each symbol. The moving action is composed of two semantic symbols, such as "Move" for moving from A to B and "Return" for coming back to me. From this assumption, the robot cannot grasp two objects simultaneously. If a robot can perform a high number of actions, there will be many corresponding action semantic symbols in the proposed method.

This brief assumption also leads the problem about the mobile robot's navigation without collision; therefore, we used a virtual environment to verify our semantic model before testing in a physical environment. Our virtual robot wandered throughout a virtual home without collision and map based path planning. Figure 2 shows the virtual environment that the robot navigated.

In this home environment, places such as a kitchen and a room are defined as elements of place in a semantic symbol group. House fixtures, such as chairs, tables, televisions, or the like are also defined as elements of place in a semantic symbol group. Further, humans in the environment, including the user have position information. Therefore, humans are included in the place semantic symbol group. For simplicity, only a two-person environment was considered in this study. Semantic symbols related to objects in the home environment are defined as object semantic symbols. The object semantic symbol group includes the object semantic symbol which a robot can grasp or press like cups, water bottle, flower, switch etc. To sum up, all the semantic symbols include the following: the action semantic symbol group, the place semantic symbol group, and the object semantic symbol group.

In fact, many more semantic symbols should be added to this system to enable an abundance of expression. However, any additional semantic symbols would be assigned to the existing symbol definition groups. For instance, "book" is not yet included in our semantic symbols, but its symbol could be assigned to an existing group. Defined groups and detailed semantic symbols, including action semantic symbols, are arranged in Table 1.

Table 1 Symbol groups and element of each group

Action	Move	Grasp	See
	Return	Release	Press
Place	Kitchen	Room	Sink
	Shelf	Table	Chair
	Refrigerator	Television	Floor
	Wastebasket	User	Another person
Object	Cup1	Cup2	Cup3
	Water bottle	Juice can	Cola bottle
	Milk bottle	Flower	Light switch
	TV switch	Apple	Banana

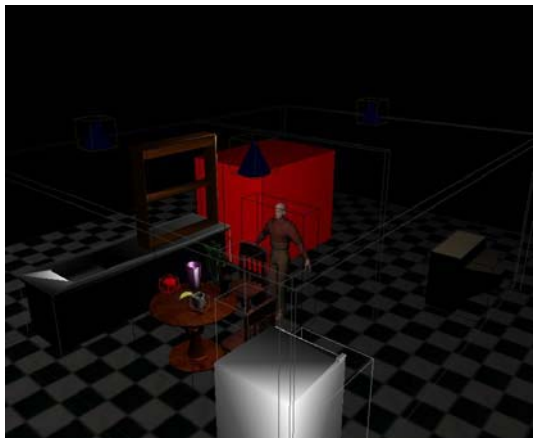

Fig. 2 Assumed environment of a robot

2.3 Tasks of robot

In a typical home environment, a user may want the robot to do many tasks. Noticing which task the user commands is the output of this system following a given input. According to the survey report of the intelligent service robot project, conducted in Sweden by Z. Kahn [7], most tasks that ordinary people want robots to do in the home environment are house chores, such as cleaning and helping with heavy loads. Such tasks are mainly "Fetch and Carry" type tasks; therefore, this study focused on "Fetch and Carry" tasks. The robots just can do moving, grasping, seeing, releasing, and pressing and it performs tasks in the home environment. If we consider the assumed environment and the basic actions that the robot can carry out, the most probable expected tasks are the following:

- Task 1: Go somewhere
- Task 2: Fetch something
- Task 3: Move something to another place
- Task 4: Show me something/someplace
- Task 5: Turn a switch
- Task 6: Take something from me and put it somewhere

To be sure, many combinations of each action could be performed, such as "Move to another place and return, move another place and return". However, these combinations do not have any meaning as a task of robot, so they are not thought over as a task. Moreover, one sentence is taken into account as a command. That is, multiple tasks, such as "Fetch a cup from kitchen, then take a water bottle from me and put it

on the table" is now not considered. We are now going on this topic.

In this section, the relation between the multimodal information and the semantic symbols is discussed as an input modality. Then, the robot environment and actions the robot is capable of are presented. Objects and some locations in the environment are expressed in a semantic symbol form, representing the "knowledge" of the robot and action semantic symbols are considered. Finally, tasks are assigned to the robot in a given environment. In the next section, we will consider problems that may occur and discuss possible solutions.

3. USER INTENTION ABSTRACTION

3.1 Probabilistic model for intention abstraction

When a semantic symbol sentence is presented to the robot, the robot should understand the intended task commanded by the user. When a user commands a task in semantic symbol, however, the following problems could occur:

1) Same meaning, different expression: People usually do not express their intention in a fixed manner. As mentioned in the introduction to this paper, users alter their same intention from time to time. If the developer of interaction channel restricts input form, users would think this system is very poor and feel uncomfortable.

2) Insufficient input information: An insufficient information problem could occur if the robot interprets information given to it as insufficient. A user may think that he or she has given sufficient information to the robot, and they would think that the interpretation of the input is remained to the robot regardless of input information. This is not a mistaken thought. Therefore, a robot should have some ability to interpret inputs that have deficient information, even if such a capability is not flawless.

Many robotics researchers have tried to solve such problems. A general approach to such problems is to make some rules: to construe the meaning of given information, some rules governing interpretation are made, and the information is then processed according to the rules. This approach is mainly used in natural language processing (NLP) research. In NLP research, to make some rules the usual grammar becomes the rules and this grammar model is constructed based on a grammatical or a probabilistic approach. Currently, this method is providing good results in some work domain. However, this approach also has some problems when information does not correspond to a given set of rules. Therefore, additional rules are needed whenever an exception is found. Further, there would be many exceptions because semantic symbol language heads for free expression rather than grammatical or formal expression. For instance, when a person wants a cup, he or she can express the command with a variety of semantic symbols:

1) Move Kitchen Grasp Cup Move Me
2) Kitchen Grasp Cup Move Me
3) Kitchen Cup Me
4) Cup

Therefore, a probabilistic approach is considered rather than a grammatical method to cover the many exceptional cases. For a probabilistic model to be considered as an interpreting input module, it should achieve the following:

1) The model (or models) can discriminate the meaning of input sentences to notice what task should be conducted. Since, specified tasks could have similar actions or semantic symbols for objects or places, the model must be able to distinguish similar input sentences.

2) The model (or models) is flexible, regardless of a given input sentence dimension. If an input sentence is considered, the dimension of the sentence cannot be estimated. That is, the total number of semantic symbols in one sentence cannot be predicted because a user may select semantic symbols arbitrary. Therefore, a model that can manage changeable input dimensions is needed.

3) The model (or models) can represent a sequence of semantic symbol inputs. For example, although the commands "Move-Return" and "Return-Move" (for simplicity, objects, or place semantic symbols are omitted) have the same semantic symbols, the sequences are different and the different sequences represent different meanings.

A Markov model can be a used to achieve these goals. A probabilistic model, the Markov model is widely used in pattern classification problems. A Markov process is a process that moves from state to state, depending on the previous n states. This model bears its meaning following reasons. First, it can distinguish the meaning of inputs. Supposing that each Markov model is assigned to each task, the transition probabilities would be different according to the characteristics of the individual task. Moreover, this model can settle input dimension problems, because endless transitions between states are possible according to model type. Finally, the sequence problem can be solved because the two transition probabilities, from A state to B state and from B state to A state, are different.

Therefore, a Markov model is constructed for our proposed method. This model is an ergodic type, which has full connections with other states, and it is a first order Markov model. The start state and end state are connected to all 10 states respectively. (Naturally, the start state and end state are not connected directly. This case cannot occur). Figure 3 shows the defined model.

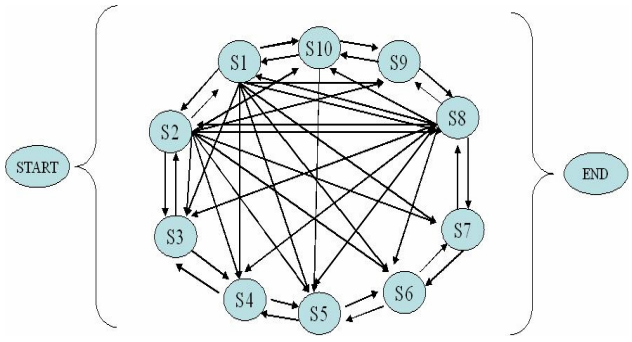

Fig. 3 Defined first order Markov model

In a first order Markov model, state transition from a state depends on the previous state alone. This is a well-known Markov assumption. This relation is expressed by equation (1) for a state sequence $\{S_1, S_2, ..., S_n\}$.

$$P(S_n \mid S_{n-1}, S_{n-2}, ..., S_1) = P(S_n \mid S_{n-1}) \quad (1)$$

Therefore, the probability of the total sequence is calculated by equation (2).

$$P(S_1, ..., S_n) = \prod_{i=1}^{n} P(S_i \mid S_{i-1}) \quad (2)$$

As mentioned in section 2 of this paper, all the semantic symbols in this method are divided into three subgroups: an action group, a place group, and an object group. The object

group is re-divided into two sub groups, according to its semantic attributes: “graspable object” or “press-able object.” Similarly, the place group is separated into two subgroups, according to the final position of the robot: “before the user” and “general locations.” By dividing a state, the ability to discriminate can be increased, but over-dividing a state may cause performance problems.

The reason why the groups are referred at this moment is that these groups are used as states of the model. With regard to a given action group, each element of the action group, such as a basic action, becomes the state of the model like {Grasp} state. Finally, start and end state is included. When counting all states, total 12 states are defined. Table 2 shows the arranged state and corresponding semantic symbol elements. For a simple description of the states, each state is numbered like S1 in Table 2.

Table 2 Defined states and their elements

State	State Attribute	Element(s)
S1	Move	Move
S2	Grasp	Grasp
S3	See	See
S4	Return	Return
S5	Release	Release
S6	Press	Press
S7	Graspable objects	Cup1, Cup2, Cup3, Water bottle, Juice can, Cola bottle, Milk bottle, Flower, Apple, Banana
S8	Press-able objects	Light Switch, TV switch
S9	Before a user	User
S10	Usual locations	Kitchen, Room, Sink, Shelf, Table, Chair, Refrigerator, Television, Floor, Wastebasket, Another Person

In a Markov model, each state corresponds to an observation. A semantic symbol in an input sentence is an observation in this case. If an observed input sequence is “Kitchen-Move,” the sequence probability is calculated using equation (2).

P(observed input sequence) =
P(PLACE|START) P(MOVE|PLACE) P(END|MOVE)

If there were only one Markov model, the discrimination of user intention would be difficult. Therefore, 6 Markov models are constructed corresponding to 6 tasks. Since 6 models are constructed, 6 result possibilities can be calculated from one input sentence. The structure of these models is shown in Fig. 4.

The transition possibilities in each model would be different, because the input patterns have different characteristics corresponding to the individual tasks. Hence, the six result probabilities are calculated in relation to the given input, and a maximum probability can be found by comparing the six calculation results. The task that has the maximum probability is determined to be the user intention. The raw user intention, which may be ambiguous, is abstracted to a task in this way. However, to guarantee the capability to discriminate, transition probabilities that encompass general input patterns are required. This issue will be discussed in the next section.

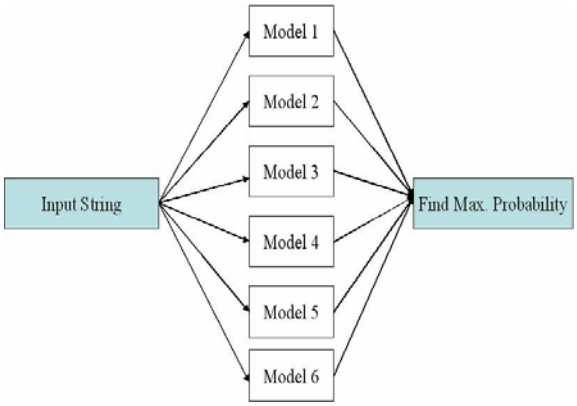

Fig 4 The structure of six models and input/output relation

3.2 Questionnaire survey

Transition probabilities encompassing many situations are necessary to achieve high classifying ability. Because the transition probabilities are independent of time basically, they should be obtained with care. To obtain transition probabilities, a questionnaire survey is conducted. By allowing prototypical users to command many specified tasks, general input patterns can be obtained. Because individual users do not express always express their task commands in the same manner, various patterns can be acquired. Less formal inputs, which may be ungrammatical, can be processed in this way.

The questionnaire survey was given to ordinary people. Because the survey takers do not usually know how to specify or command an input precisely, four or five example input commands for each task are provided as guides, though various expressions of inputs are encouraged. However, to avoid a simple translation from the questionnaire example sentences, written in usual language to semantic symbol language, specific circumstances are given to the subjects and it is assumed that the position of objects or locations in the environment is known by both a user and the robot. The following are some examples for task 2, which specifies “Fetch something”:

- You are now thirsty. You want to make the robot fetch something to drink for you.
- You want to make the robot bring you a flower on the table.
- You want to eat an apple that is on the sink. Make the robot fetch it for you.

In all, 56 subjects participated in the questionnaire survey. 71% of the subjects were male participants, and 29% were female. The average age of the subjects was 30.9 years. The ages of the participants were from 20 to 59, and the age group of 25 to 29 was the largest, at 41.1% of the participants. 33.9% of the participants were employed and 66.1% were students. Approximately 175 command samples for each task were obtained, so 1051 command samples were obtained for the 6 tasks. From the results of the survey, 6 transition probability tables were prepared. Table 3 shows a section of the probability table for task 3 as an example.

Table 3 Transition probabilities corresponding model 3

i-1 \ i	S1	S2	S3	S4	....	END
START	0.5359	0.0497	0.0276	0	...	0
S1	0	0.0160	0	0	...	0.0562
S2	0.2531	0	0	0	...	0.0189
S3	0.1666	0	0	0	...	0
S4	0	0	0	0	...	1
S5	0.0070	0	0	0.0352	...	0.8521
S6	0	0.5	0	0	...	0
....	...	...	...	...	...	...
S10	1	0	0	0	...	0

4. EVALUATION

An evaluation is conducted to check how well the robot maps the commands specified by the users and the interpreted results by the robot. This evaluation can be presented clearly using a confusion matrix. Table 4, 5 shows the results. The upper right row represents the commands that are understood by the robot and the left column represents the user commands. A separate confusion matrix is constructed to verify whether there is a difference between the commands that are included in the probability tables and commands that are not included, but Table 5 does not have the results of task 6 because example questionnaire about task 6 are not presented to the user at first. Collecting more information about task 6 will be performed.

Table 4 Confusion matrix: the test commands are included in the transition probabilities

		Interpreted result					
		Task1	Task2	Task3	Task4	Task5	Task6
Actual User Command	Task1	74	5		1		
	Task2		74	5			1
	Task3		1	74			5
	Task4	5			75		
	Task5					80	
	Task6		2				78

Table 5 Confusion matrix: the test commands are not included in the transition probabilities

		Interpreted result					
		Task1	Task2	Task3	Task4	Task5	Task6
Actual User Command	Task1	18					
	Task2		17		1		
	Task3		2	16			
	Task4	2			16		
	Task5	3	1			14	
	Task6						

Tables 4 and 5 show that user commands were well interpreted to some degree. Though these results were obtained from only a few test commands, certain tendencies can be observed from them. The obtained results show that the proposed method has a promising ability to interpret a user's intention. For example, when the task 6 command, "Receive water bottle from me, and put it on the shelf," is expressed as "Grasp-Shelf-Release," the interpreting part made right answer. Similarly, when the "Move-See-Water bottle-Return" is given for the task 2 command, "Fetch the water bottle," the interpreted result is also correct. However, the incorrect results occurred when the user commands were too abstract, that is, the information presented is insufficient, such as "Grasp Cola" for the task of "Fetch the cola." In this case, even a human could not interpret the sentence at one time.

Another problem is that the transition probability tables do not contain all the cases. Although the transition probability tables include general information corresponding to tasks, they are not useful in some special cases. However, many situations can be distinguished by multiplying the transition probabilities sequentially, and there are many cases for which the input sentence does not make sense when a very special input is presented.

5. CONCLUSIONS

This paper presents a method to enhance human-robot interaction, by which a multimodal information source is a human user's voice command or visual information about objects, places, and actions. This multimodal information is represented as semantic symbols, and these semantic symbols each have word-level meaning. Some robot actions are defined, and the robot actions are also expressed as semantic symbols through voice modality or by gestures. The robot tasks were also discussed. In a domestic environment, probable tasks are selected and on choosing them the capability of a robot is considered. In this study, six tasks were selected to evaluate the method; the selected tasks were mainly related to fetch and carry tasks, or simple errands.

After all, users command their input through a semantic symbol array. As a pre-user intention processor, the interpreting module starts to interpret the input sentence. In other words, the robot tries to understand what task is specified by the user. To select the most probable task, six probabilistic models are constructed. These models are first order Markov models, which have several advantages in manipulating semantic symbol arrays. They can process a disordered input sentence, or one that has insufficient information, to help the robot identify which task to perform. The states of the model are defined, according to the semantic attributes of the input symbols. A total of 12 states, including the start and the end state, are selected.

The transition probabilities in each model reflect the characteristics of each task. To obtain the transitions, a questionnaire survey was conducted. By making ordinary people command the specified tasks, many command patterns were acquired. Various input patterns could be stored. Accordingly, the interpreting module could have the capability to process an unexpected input sentence. Using the obtained transition probabilities, six result probabilities corresponding to six tasks can be calculated by the consecutive multiplication of the transition probability. Because the models have different characteristics that correspond to individual tasks, the six result probabilities have different values. By comparing these resultant probability values, a maximum probability can be found. Ultimately, the task that has a maximum value is considered the user-specified task. The evaluation of the proposed method indicated that this method has good potential for user intention abstraction.

While the interpreting part is designed under the assumption that conversion of multimodal information, such as voice commands and vision information has already been implemented, an actual experiment confirming this assumption has not yet been conducted. Further experiments will be conducted to verify whether our method is practical. In

addition, data about user commands are being collected continuously for more general transition probabilities.

ACKNOWLEDGMENTS

This paper was performed for Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Science and Technology of Korea. The authors would like to thank Jung-Yean Yang, Soo-Chul Lim, Hyoung-Rock Kim, co-workers in our team, for their useful comments on this research and we gratefully acknowledge the contribution of many survey subjects.

REFERENCES

[1] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor, and B. Charlie, "MATCH: An architecture for multimodal dialogue systems," *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 376-383, 2002.

[2] M. Johnston, "Unification based multimodal parsing," *Proc. of the 17th International Conf. on Computational Linguistics,* pp. 624-630, 1998.

[3] S. L. Oviatt, P.R. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, and D. Ferro, "Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions, Human Computer Interaction," *Human Computer Interaction*, Vol.15, No. 4, pp. 263-322, 2000.

[4] R. Bischoff and V. Graefe, "Dependable Multimodal Communication and Interaction with Robotic Assistants," *IEEE International Workshop on Robot and Human Interactive Communication ROMAN200*, pp.300-305, 2002.

[5] R.A. Bolt, "Put that there: and gesture at the graphics interface," *Computer Graphics*, Vol.14, No.3, pp262-270.

[6] D. Perzanowski, A. C. Schultz, W. Adams, E. Marsh, and M. Bugajska, "Building a Multimodal Human-Robot Interface," *IEEE Intelligent Systems*, Vol.16, No.1, pp16-21, 2001

[7] Z. Kahn, "Attitudes towards intelligent service robots," Iplab, Nada, Royal Institute of Technology.