Reading a Robot's Mind: A Model of Utterance Understanding based on the Theory of Mind Mechanism

Tetsuo Ono **Michita Imai**
ATR Media Integration & Communications Research Laboratories
2-2 Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0288 JAPAN
{tono, michita}@mic.atr.co.jp

Abstract

The purpose of this paper is to construct a methodology for smooth communications between humans and robots. Here, focus is on a *mindreading* mechanism, which is indispensable in human-human communications. We propose a model of utterance understanding based on this mechanism. Concretely speaking, we apply the model of a mindreading system (Baron-Cohen 1996) to a model of human-robot communications. Moreover, we implement a robot interface system that applies our proposed model. Psychological experiments were carried out to explore the validity of the following hypothesis: by reading a robot's mind, a human can estimate the robot's intention with ease, and, moreover, the person can even understand the robot's unclear utterances made by synthesized speech sounds. The results of the experiments statistically supported our hypothesis.

Introduction

In our everyday communications, we unconsciously attempt to read the minds of other people while trying to understand what they are saying. That is, *mindreading* is a daily activity of humans used to estimate the mental states of others by means of observing their behaviors. Recently, research has been actively carried out on such mindreading in communications due to the recognition of its importance (Premack & Dasser 1991; Baron-Cohen 1996). In addition, the general framework for such research is called "Theory of Mind."

We show concretely that the ability of mindreading is essential for a social being to communicate with others. For example, let us assume that A and B are two persons. Person A, who is carrying some bags in his arms, turns his eyes to B and utters words that are unclear speech sounds. B also turns his eyes to A, meeting A's eyes. Afterwards, A turns his eyes in the direction of his own movement. As B turns his eyes in the same direction simultaneously, he detects an object. B then recognizes the object as an obstacle for A by reading his own mental states of desire and goal. Finally, B understands A's utterance despite his unclear speech sounds and subsequently removes the object. As mentioned above, we can daily observe such reciprocal acts when communicating with others by the reading of intentions, desires, and goals.

Even in communications between humans and robots, such mindreading is indispensable for reciprocal acts. This is because we often judge synthesized speech sounds from artifacts to be strange, and we occasionally find the sounds hard to understand. We will not be able to communicate with robots smoothly as long as we do not make the reasons clear. If we do not assume the reciprocal acts mentioned above, it is possible to model communications as a "code model." A code model is a framework of signal transmission where a sender gives information (signals) to a receiver using a presupposed common code for encoding and decoding and then alternates turns within a given time. A code model, however, is unable to grasp essential qualities in communications involving humans. This is because humans cannot jointly own "code" with other entities in principle (Clark & Marshall 1981). Consequently, communications with humans cannot be grasped by transmitting restricted signal patterns.

To overcome the problem mentioned above, Sperber (Sperber & Wilson 1986) proposed bringing an "inference" viewpoint to the process of communications. In the relevance theory he proposed, humans communicate among themselves by inferring the minds of others. Certainly, this theory might approach the essence of human communications in overcoming the bottleneck of the code model. However, inference in the relevance theory can only be executed by mere "deductive" inference rules (Kimura 1997); therefore, this theory is equivalent to the code model if we regard the inference "rules" as a complicated "code." Moreover, this theory has the problem of technical terms in the theory not being connected through physical existence. These unsolved problems lower the theory's ability to explain actual communications.

The purpose of this paper is to construct a methodology for smooth communications between humans and robots by focusing on the mindreading mechanism. Es-

pecially, we consider a mechanism of utterance understanding focusing on *sympathetic* and *embodied* inference and a viewpoint from an *internal observer* in communications. We first propose a model of utterance understanding based on the mindreading mechanism. Concretely speaking, we apply the model of a mindreading system (Baron-Cohen 1996) to a model of human-robot communications. Next, we implement a robot interface system that applies our proposed model. Third, we conduct psychological experiments to explore the validity of the following hypothesis: **By reading a robot's mind, a human can estimate the robot's intention with ease, and, moreover, the person can even understand the robot's unclear utterances made by synthesized speech sounds**. Finally, we discuss the validity of our model on the basis of the results of the experiments and conclude the paper.

Understanding a Robot's Utterance

In this chapter, we propose a model of utterance understanding and a robot interface model based on a mindreading mechanism.

"Mindreading System" and Model of Utterance Understanding

In this section, we propose a Model of Utterance Understanding based on the Theory of Mind Mechanism (MUUToMM). The left-hand side of Figure 1 shows an outline of MUUToMM. Concerning the theory of mind, Baron-Cohen is constructing the most detailed theoretical model known at present (Baron-Cohen 1996). This model of a *mindreading system* assumes the following four modules relevant to each other but functionally independently.

- Intentionality Detector (**ID**): This module is a perceptual device that interprets motion stimuli in terms of the primitive volitional mental states of goal and desire.
- Eye-Direction Detector (**EDD**): This module detects the presence of eyes and computes where the eyes are directed.
- Shared Attention Mechanism (**SAM**): This module builds triadic representations that specify that Self and Other are both attending to the same Object.
- Theory of Mind Mechanism (**ToMM**): This module infers the full range of mental states from behaviors.

Among these modules, **ID** and **EDD** are modules concerning dyadic relations, while **SAM** and **ToMM** are modules concerning triadic (or more) relations. These modules perform an important role in reciprocal acts, that is, reading the minds of others in communications.

In our proposed model MUUToMM, the following module functions by the activation of the above modules.

- Utterance Understanding Mechanism (**UUM**): This module is a system for understanding the utterances of others by dynamically interacting with the process of inference in **ToMM**.

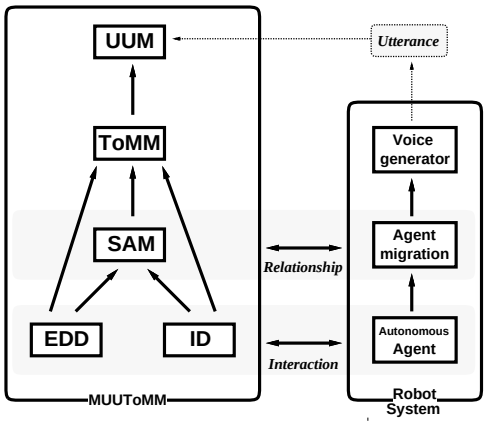

Figure 1: Outline of a model of utterance understanding and corresponding robot system.

Here, the process of understanding utterances in MUUToMM is carried out for the example described in the Introduction. Person A, who carries some bags in his arms, turns his eyes to B and utters words that are unclear speech sounds. B recognizes A as a *person* through the activation of **ID** and **EDD** by observing his behaviors mentioned above. After their eyes meet, A turns his eyes in the direction of his own movement. As B also turns his eyes in the same direction, he detects an object. This is the process of a triadic relation being constructed among A, B, and the object, which **SAM** enables them to construct. B, moreover, recognizes the object as an obstacle for A by using the function of **ToMM**. Finally, B comes to understand A's utterance by interacting with the estimated intention of A and deficient information in his utterance (**UUM**) and subsequently removes the object.

Our proposed model enables us to overcome the problems in the "code model" and the relevance theory. This is because our model considers a mechanism of utterance understanding in connection with *physical existence* and regards the *relationship* that emerges between a speaker and a hearer as important. In other words, our model can clarify the mechanism of human communications in the real world which cannot be completely reduced to "symbolic" inference. In the next chapter, we explore the validity of this model through psychological experiments.

Robot Interface System

In this section, we describe a robot interface system that applies the model of utterance understanding MUUToMM (the right-hand side of Figure 1). We use robots because we consider communications in the physical world where humans live and also because robots enable us to control parameters in experiments. A serious problem in human-robot communications is that humans and robots cannot construct relationships sim-

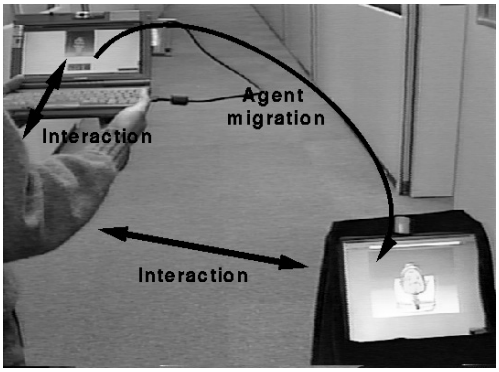

Figure 2: Outline of robot interface system.

ilar to human-human ones, since humans do not regard robots as autonomous beings with intention (the human **ID** does not function) and do not call their attention (the human **SAM** does not function).

In our robot interface system, the functions of human **ID** and **EDD** are activated by an autonomous agent, and the function of **SAM** is activated by a mechanism of its agent migration (the shaded area of Figure 1 shows the corresponding relations). Figure 2 shows the implemented robot system. The following definitions are applied in our system.

Autonomous agent: An agent in our system can behave autonomously because it adopts a model of Multi-Order Functions (MOF) (Ono & Okada 1998) as its internal model. We can observe both predictable and unpredictable behaviors of the agent because the MOF model spontaneously changes and consistently maintains the internal states while interacting with the environment. Accordingly, since each agent with the MOF model behaves autonomously while interacting with humans, the human **ID** and **EDD** become activated. The interaction in Figure 2 is similar to the interaction between a user and a digital pet.

Agent migration mechanism: In our system, a mechanism of agent migration (Ono, Imai, & Etani 1999) enables the attention of a dyadic relation to be moved to a triadic one. Concretely speaking, an agent can migrate from a user's mobile PC to a robot. As a result of this migration, the robot can inherit the user's attention from the agent, enabling a relationship to form among the user, the robot, and an object along the direction of the robot's own movement (the human **SAM** becomes activated; Figure 2).

Voice generator: The robot in our system can make utterances with synthesized speech sounds. Because synthesized sounds are used, we can freely set up parameters in the utterances such as intonation, accent, and clarity.

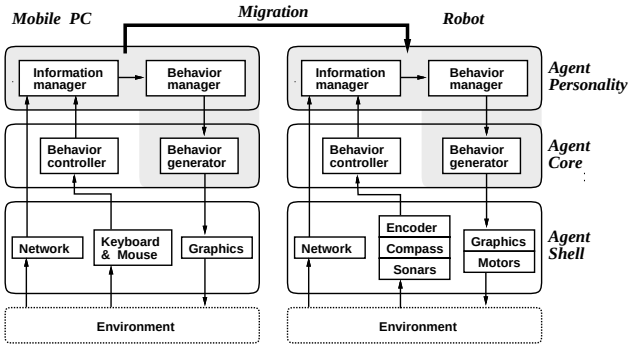

Figure 3: Structure of agent and mechanism for agent migration.

Implementation of Agent and Migration Mechanism

In this section, we briefly describe the structure of the agent and the mechanism for agent migration (Figure 3). The life-like agent consists of three layered components: an agent personality (AP), an agent core (AC), and an agent shell (AS), which can be rearranged dynamically. The AP has knowledge-based objects related to the user and environment, the AC has processing definitions for path-planning and behaviors, and the AS controls the physical resources of the network and the robot. In the process of migration, the AP and part of the AC, i.e., the shaded area of Figure 3, move from the mobile PC to the robot. An unpossessed robot can only move autonomously while obeying the set initial state.

In our experiments, we temporarily simplify the system setup, the interaction with the user, and the mechanisms of the agent and robot. This is because the aim of these experiments is basically to test whether subjects can understand the utterances of the robot under changing conditions and parameters. The interaction is like that between a user and a digital pet. The user first gives the agent a stimulus by clicking a mouse so that the agent changes its internal states. The agent changes the states by itself with a mechanism to generate autonomous and multiple behaviors (Ono & Okada 1998). Similarly, the agent's migration is simplified: the agent migrates from the mobile PC to the robot automatically under the experimental condition when the robot stops in front of an obstacle.

Experiments

In this chapter, we conduct experiments to test the validity of the proposed model of utterance understanding by using the implemented robot system.

Method

The experiments were conducted by the following method.

Subjects: Twenty-seven undergraduate and graduate students (male and female). The subjects were randomly divided into three groups: seven subjects were assigned for a preliminary experiment, ten for an experimental group, and ten for a control group.

Environment: Figure 4 shows an outline of the experimental setup. Points A-E in Figure 4 denote the positions of a robot, a subject, an experimenter, and an observer. The arrow from A to B is a trace of the robot. The subject's behavior is observed through a camera.

Preliminary experiment: For seven subjects, we examined their levels of understanding for three utterances made by synthesized speech sounds with changing sound parameters. We finally adopted one utterance for the experiments, which three subjects out of the seven understood. The content of the utterance is "Move the trash can out of my way."

Conditions: We prepared two conditions, which differed in their processes of interaction (*Interaction factor*). Under the experimental condition, the "character"[1] migrated from a mobile PC to a robot. Under the control condition, the character did not migrate. The subjects were distributed among the two conditions randomly. Moreover, the subjects psychologically evaluated the character and the robot by completing questionnaires (*Target factor*).

Procedure: The experiments consisted of the following four phases.

1. The subjects received the following instructions from the experimenter (position C): "This experiment is part of research concerning character design." The subjects, moreover, were taught a method to interact with the character on the mobile PC, and they actually practiced the operation for five minutes. After that, they psychologically evaluated the character through the questionnaires.
2. The experimenter told the subjects that he forgot to take a tool, and he moved from C to D in Figure 4.
3. Three minutes after leaving the experimenter, the robot approached the subject from A and stopped in front of the trash can (before that, the subject was unaware of the robot). Then, **under the experimental condition, the character migrated to the robot; meanwhile, under the control condition, it did not migrate**. The robot, moreover, gave the utterance that was adopted in the preliminary experiment. The content of the utterance was "Move the trash can out of my way."

[1]The word "agent" has various meanings in various fields. Accordingly, the more general word "character" is used in these experiments in consideration of the influence on the subjects.

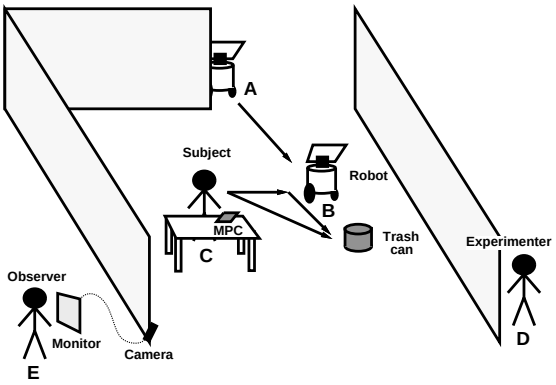

Figure 4: Experimental setup: arrangement of subject, observer, and robot.

4. The subjects psychologically evaluated the robot through the questionnaires when the experimenter came back to C two minutes later.

Evaluations: The results of the experiments were evaluated from the answers of the questionnaires and the record of the subjects' behaviors. In the questionnaires, the subjects were asked whether they understood the robot's utterance, why they behaved the way they did toward the robot, and to give a psychological evaluation of the character and the robot. In the evaluation, the subjects provided answers on a ten-point scale for two items: *Autonomy* and *Familiarity*, which are composed of the following three subitems respectively. The average score of each subitem was used in analyzing the results.

- Autonomy: Motivated ↔ Spiritless, Active ↔ Passive, Free ↔ Restricted
- Familiarity: Familiar ↔ Not familiar, Pleasant ↔ Trying, Kind ↔ Unkind

Hypothesis and Predictions

We put forward the following hypothesis: **By reading the mind of the robot to whom the agent migrated from the mobile PC, subjects under the experimental condition can estimate the robot's intention with ease. Moreover, the subjects can even understand the robot's unclear utterances made by synthesized speech sounds**. In the experiments, we aimed to verify the following three predictions derived from the hypothesis. Under the experimental condition compared with the control one,

Prediction 1: The subjects will regard the robot as an autonomous entity with intention (the function of **ID** and **EDD**).

Prediction 2: The subjects will first look at the robot and then turn their eyes to the trash can (the function of **SAM**).

Prediction 3: The subjects will be able to estimate the robot's intention with ease, and this will facili-

Table 1: Number of subjects who understood robot's utterance.

	Understanding	No-Understanding
Control	3	7
Experimental	8	2

Table 2: Number of subjects acting on robot's command.

	Acting	No-Acting
Control	1	9
Experimental	8	2

tate their understanding of the robot's utterance (the functions of **ToMM** and **UUM**).

Results of Experiments

In this section, we verify the three predictions on the basis of the results of the experiments. We verify them in the inverse order from prediction 3 to 1 to make the point of the argument clearer.

Verification of Prediction 3 First of all, we verified Prediction 3. In this experiment, we asked the subjects in the questionnaires whether they understood the robot's utterance. Table 1 shows the results for the number of subjects who understood it and those who did not under both conditions. In the analysis, a significant difference was found ($\chi^2 = 5.051, p < .05$). Accordingly, the difference between conditions (*Interaction factor*) had an effect on the subjects' understanding. In other words, **although the subjects understood to a large extent the utterance of the robot to which the agent migrated (Experimental condition), the subjects did not understand very well that of the robot to which the agent did not migrate (Control condition)**. To support this observation, almost all of the subjects who satisfied the robot's request were in the experimental group (Table 2). Figure 5 shows the appearance of a subject satisfying the robot's request, "Move the trash can out of my way." In contrast, Figure 6 shows the appearance of a subject not understanding the request. The difference between both conditions is caused by the agent migration because the utterance that all of the subjects received was the same synthesized speech sound utterance.

Verification of Prediction 2 Next, we verified Prediction 2. From the results of observations on the subjects' behaviors, we calculated the average time that they fixed their eyes on the robot and the trash can. From the results of the calculations, the average time of the experimental group was 33.7 seconds; in contrast, that of the control group was 28.1. However, we were unable to compare the times of both conditions simply because the time in the experimental group was only counted until the trash can was removed. Moreover, we could not judge distinctly whether the subjects fixed their eyes on the robot or the trash can because we did not use a tracking device like an eye-camera.

Figure 5: Photo of subject understanding robot's utterance.

Figure 6: Photo of subject not understanding robot's utterance.

Even considering these problems, however, we obtained evidence from the remarks in the questionnaires that all of the subject under the experimental condition noticed the trash can; however, half of the subjects under the control condition did not notice it. Accordingly, the former subjects turned their eyes from the robot to the trash can along the robot's running direction as the agent migrated; however, the latter subjects did not turn their eyes. It can be concluded that the agent migration prompted the subjects to turn their eyes because the only difference between the two conditions was this agent migration.

Verification of Prediction 1 Finally, we verified Prediction 1. We analyzed the results of the questionnaires to test whether the subjects regarded the robot as an autonomous entity with intention. In the analysis, we estimated the subjects' impressions of the robot by using two evaluation items, i.e., *Autonomy* and *Familiarity*, and their remarks. First, we tested whether the individual scores of the items in the questionnaires exhibited crossover interaction between the *Interaction factor* and *Target factor*. Consequently, crossover interaction was exhibited for *Autonomy* ($F(1, 18) = 14.223, p < .01$; Figure 7); additionally, a main effect was exhibited for the *Target factor*. Furthermore, crossover interaction was exhibited for *Familiarity* as well ($F(1, 18) = 9.31, p < .01$; Figure

Table 3: Summary of experimental results.

	Experimental	Control
Understanding the robot's utterance	Yes	No
Focusing on the trash box	Yes	Nearly all No
Considering the robot's intention	Yes	No

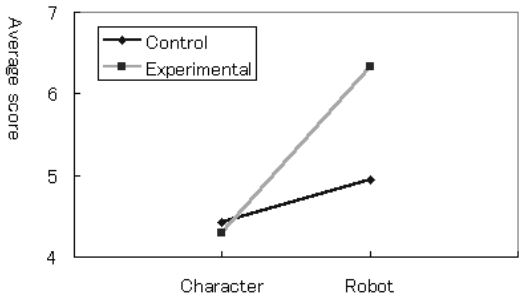

Figure 7: Interaction between two factors on *Autonomy*.

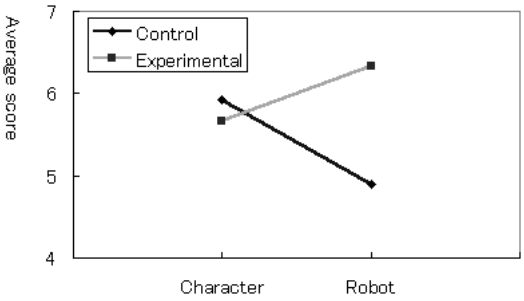

Figure 8: Interaction between two factors on *Familiarity*.

8); however, no main effect was exhibited.

In summary, as a result of migration, a robot can inherit the relationship from the interaction between a subject and an agent so that a relationship is formed between the subject and the robot. The questionnaires also asked the subjects who satisfied the robot's request why they did so. The subjects answered as follows: "I felt pity for the robot not going forward" and "The robot felt like going forward." As a result, the subjects in the experimental group regarded the robot as an autonomous and familiar entity, and looked at the robot as if it had an intention and an aim.

Summary of Experimental Results Table 3 shows a summary of the experimental results. Based on the consideration mentioned above, the three predictions were verified. Consequently, the hypothesis was supported by the experiments.

A noteworthy point in the experiment is that all of the subjects under both conditions turned their eyes and paid attention to the robot (in Verification of Prediction 2). As a result of their behaviors, although they should detect the trash can beside the robot's, half of the subjects under the control condition did not notice it. In other words, although all of the subjects could get the same contextual information, there were considerable differences in their understanding of utterances and behaviors between both conditions. We can attribute these differences to whether a mechanism of mindreading is activated. These differences cannot be fully explained only by a function of attention and symbolically situated inference. In the next chapter we discuss this issue.

Discussion and Conclusions

The purpose of this paper was to construct a methodology for smooth communications between humans and robots. In particular, we focused on the mindreading mechanism that is indispensable in human-human communications. Moreover, we proposed a model of utterance understanding based on this mechanism and implemented a robot interface system that applied our proposed model. In psychological experiments with the implemented system, the subjects in the experimental group activated "universal modules" (**ID** and **EDD**) in perceptions of the robot because an agent interacting with the subjects migrated from their mobile PC to the robot. Afterwards, a triadic relationship among the subject, the robot and an obstacle was constructed (**SAM**). On the basis of this relationship, by reading the robot's mental states of desire and goal (**ToMM**), the subjects could even understand the robot's unclear utterances (**UUM**). However, the subjects in the control group could not understand the utterances and construct the relationship despite getting the same contextual information as the subjects in the experimental group. These results are not explained only by a function of attention and symbolically situated inference. This is because all the subjects paid "attention" to the robot; moreover, there was no difference between the experimental situations as far as we could describe them through "symbolic" representation.

We can attribute the subjects' understanding of the robot's utterance to the activation of a mechanism of mindreading. We also believe that this activation is caused by the interaction among the five modules in our model (Figure 1). However, as this model still has a few shortcomings, we must further investigate the mechanism in detail. With the progress of this research, the following concepts become important: *sympathetic* and *embodied* inference as well as a viewpoint from an *internal observer* in communications. An impasse in the research on the theory of mind resulted from adopting tasks in experiments that subjects could solve by symbolic inference alone, e.g., "false belief task" (Wimmer & Perner 1983). Owing to adopting the task, an

autistic child who is considered not able to estimate others' mental states can manage to carry out the task by repeated training. This is because he/she has a sufficient ability to think with operating symbols. However, the essential problems are that they lack the ability of sympathetic understanding to others and cannot take up the subjective attitude involved in communications. These two points are indispensable for investigating a mechanism of human communications.

In our experiment, we did research on human-robot communications from the two viewpoints mentioned above. That is to say, we could stimulate the subjects to promote the sympathetic understanding by the experimental setup where the robot as physical existence could not go forward if it was blocked by an obstacle. We could, moreover, make the subjects participate in communications with the robot subjectively by the relationship that emerged from the agent migration. As a results of the experimental setup, the subjects could even understand the robot's unclear utterances. Therefore, we could demonstrate the importance of *sympathetic* and *embodied* inference and a viewpoint from an *internal observer* in communications through the experiments.

Our proposed model can also give suggestions for problems in the relevance theory. As mentioned above, we considered the mechanism of utterance understanding focusing on the two viewpoints. Moreover, we emphasized that the former was caused by being a similar physical existence, and the latter was made possible by the relationship that emerged between a speaker and a hearer. The relevance theory is unable to overcome the code model because it is unable to adopt the above two points. We take a constructive approach to resolve these problems in human communications by regarding the above two viewpoints as important. Research on robots is particularly important in this field because it logically shifts our concerns from toy problems to real world ones and gives us empirical data through experiments on changing parameters.

Finally, we describe unsolved problems and future works. First, we need to design and construct a more effective human-robot interface for smooth communications. In our proposed model, we constructed a relationship between a human and a robot using an autonomous agent and the mechanism of agent migration. However, as a matter of course, we have to design it more naturally in other ways. For example, we can use functions found in daily behaviors, e.g., the effect of a greeting or falling into step with others. In the future, we plan to study basic factors in human-robot communications found in everyday interactions. Next, we have to consider engineering applications of our research. In our present-day society, there are many artifacts around us, such as cellular phones and computers, that can enhance our cognitive functions. We can expect robots to provide daily support to humans in the near future, e.g., carrying things, rescuing victims and guiding people, with the advance of robotics technology. Accordingly, it is important that humans can communicate with robots smoothly and reliably. We believe that our research can contribute to this field of engineering applications as well as to scientific interests in the mechanism of human communications.

References

Baron-Cohen, S. 1996. *Mindblindness.* MIT Press.

Clark, H. H., and Marshall, C. R. 1981. Definite Reference and Mutual Knowledge. In Joshi, A. K.; Webber, B. L.; and Sag, I. A., eds., *Elements of Discourse Understanding.* Cambridge University Press.

Kimura, D. 1997. Information, Regularity, and Communications – Comparison betwen Shannon and Bateson –. In Tani, Y., ed., *Communication no Shizen-shi (in Japanese).* Tokyo: Shin-yousha. 31–60.

Ono, T., and Okada, M. 1998. Consistency Generation dependent on Situation. In *7th IEEE International Workshop on Robot and Human Communication (ROMAN'98)*, volume 1, 40–45.

Ono, T.; Imai, M.; and Etani, T. 1999. Robots as Human Peers: Cognitive Conditions of Human-Robot Interaction. In *The Second International Conference on Cognitive Science (ICCS'99)*, 693–696.

Premack, D., and Dasser, V. 1991. Theory of Mind in Apes and Children. In Whiten, A., ed., *Natural Theories of Mind.* Blackwell.

Sperber, D., and Wilson, D. 1986. *Relevance: Communication and Cognition.* Oxford: Basil Blackwell.

Wimmer, H., and Perner, J. 1983. Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding Deception. *Cognition* 13:103–128.