Physical Relation and Expression: Joint Attention for Human–Robot Interaction

Michita Imai, Tetsuo Ono, and Hiroshi Ishiguro

Abstract—**This paper investigates situated utterance generation in human–robot interaction. In addition, we study the achievement of joint attention because a person must have joint attention with a robot to identify the object indicated by a situated utterance description generated by the robot. This paper proposes a joint attention mechanism to achieve such joint attention as well as a speech generation system named Linta-III. By using the joint attention mechanism, Linta-III can omit obvious information in the situation from an utterance description. The joint attention mechanism employs eye contact and attention expression functions. These functions are the robot's physical expressions, and they allow the joint attention mechanism to draw the person's attention to the same sensor information as that noticed by the robot. We have also conducted a psychological experiment to evaluate the joint attention mechanism. The results indicated that the eye contact and attention expression functions were effective methods in the development of joint attention.**

Index Terms—**Artificial intelligence, intelligent robots, joint attention, man–machine systems, oral communication, situated utterance, speech intelligibility, user interface human factors.**

I. Introduction

NEW TYPES of robots, such as pet robots, are being developed to communicate with humans. In the near future, people can expect to be in situations where they have to communicate with robots appearing in front of them. To support human–robot interaction in such situations, this paper describes a speech generation system that refers to sensor information.

People commonly refer to nearby situations in face-to-face communications. At such times, they omit obvious words to use efficient descriptions in utterances [1]. Since robots will also come to share real-world situations with people, the dialogue systems for robots must also be able to refer to nearby situations to generate or interpret utterance descriptions. In short, robots must be able to obtain sensor information relevant to the current context to deal with situated utterance descriptions.

Manuscript received December 19, 2001; revised July 31, 2002. Abstract published on the Internet May 26, 2003. This work was supported in part by the Telecommunications Advancement Organization of Japan. This paper was presented at the 2001 IEEE International Workshop on Robot and Human Interactive Communication, Bordeaux and Paris, France, September 18–21.

M. Imai was with the ATR Media Information Science Research Laboratories, Kyoto 619-0288, Japan. He is now with the Faculty of Science and Technology, Keio University, Yokohama 223-8522, Japan (e-mail: michita@ics.keio.ac.jp).

T. Ono was with the ATR Media Information Science Research Laboratories, Kyoto 619-0288, Japan. He is now with the Future University, Hakodate 041-8655, Japan

H. Ishiguro was with the ATR Media Information Science Research Laboratories, Kyoto 619-0288, Japan. He is now with the Department of Adaptive Machine Systems, Osaka University, Osaka 565-0871, Japan.

Digital Object Identifier 10.1109/TIE.2003.814769

Dialogue systems for robots have been developed with attention mechanisms [2], [3]. Such attention mechanisms focus on relevant sensors to comprehend a situation. For example, the dialogue system can interpret a situated utterance description if the attention mechanism focuses on the robot's front sensors to immediately locate obstacles as the robot proceeds forward. If a person says "No!" to the robot when it moves forward toward an obstacle, how can the robot interpret the person's word with the attention mechanism? Since the attention mechanism focuses on the front sensors, the dialogue system within the robot is able to interpret the word "No" as "do not proceed forward" based on reference to the focused front sensors. The focus of attention method is also employed in dialogue systems for computer graphics (CG) [4], [5]. The focus of attention method to CG objects reduces the computational cost when an utterance is generated because it restricts the contents of the utterance according to the relevant object. The relevance theory [6] also proposes a communication model based on recognized situations and the experiences of people. It employs the technical term "mutual manifestness," which is a mental state where two or more persons recognize the same situation or recall similar experiences. The relevance theory regards a person's communication as the process of gaining mutual manifestations by passing messages to others. The concept of mutual manifestations is similar to that of the focus of attention method in terms of using a situation.

However, the focus of attention method is insufficient for achieving the use of a situated utterance description in a speech generation system. The following three difficulties must be overcome to achieve such a system:

1) how to draw a person's attention to the information to which a robot is paying attention;
2) how to make a person understand the communicative intention of a robot;
3) how to deal with a person's attention in the attention mechanism.

Difficulty 1) is attributed to the lack of a robot's expressive ability when ordinary attention mechanisms select relevant sensors. Because of this lack, a person cannot realize where a robot is focusing its attention. That is to say, in terms of the relevance theory, the person and the robot cannot be in a state of mutual manifestation. The target of a person's attention in a conversation is a crucial problem that needs to be addressed for the robot's speech generation system to be able to deal with real-world situations. For example, when a guide robot in a museum focuses on an artwork and begins to explain it to a person by using demonstrative pronouns, the person will be unable to understand just what item the robot is explaining if the person is not focusing on that work.

Difficulty 2) derives from Difficulty 1). The relevance theory insists that the occurrence of a state of mutual manifestation depends on the listener inferring the speaker's communicative intention. For example, when a robot (i.e., speaker) says "take this away!" in front of a box (in order to proceed forward), a person must guess the robot's communicative intention (that it wants to proceed forward) to interpret the robot's utterance. If, according to the guess, the person's attention is drawn to the box, he or she can be said to have a state of mutual manifestation with the robot. Accordingly, since the listener's inference of the communicative intention has an effect on drawing a person's attention, we must deal with Difficulty 2).

Finally, Difficulty 3) is attributed to the type of focal attention achieved by conventional attention mechanisms. The conventional type includes only sensor information, not a person's attention. However, the speech generation system must also follow the person's attention to be able to generate a situated utterance description, because what the person pays attention to determines whether the person can interpret an utterance description if obvious words in the situation are omitted.

This paper proposes a joint attention mechanism that achieves joint attention [7], [8] between a person and a robot, i.e., they come to refer to the same real-world situation. Since the person has joint attention with the robot, the robot is able to generate a situated utterance.

To provide a solution for Difficulty 1), the joint attention mechanism employs an attention expression function. This manifests the sensor information under focus with a physical expression. The physical expression function employs gaze motions to face toward the information under focus and hand gestures to point to it.

To provide a solution for Difficulty 2), the joint attention mechanism employs an eye-contact function between the person and the robot to promote the relationship between them. The eye-contact function works by turning the robot's head to face the person. The resulting relationship with the robot encourage the person to guess the robot's communicative intention and to become aware of the robot's attention manifested by the attention expression.

To provide a solution for Difficulty 3), the joint attention mechanism uses an attention coordinate that represents the source direction of the sensor information focused on by both the person and the robot. The representation of the person's attention varies depending on the development of the joint attention. Since the attention coordinate reflects the person's attention, the joint attention mechanism can adapt an utterance description to suit a situation depending on the attention coordinate.

In addition, we have developed a speech generation system named Linta-III with the joint attention mechanism on a humanoid robot called Robovie.

The remainder of the paper is organized as follows. Section II explains problems in generating a situated utterance and also gives an overview of Robovie. Section III explains Linta-III and the process used by the joint attention mechanism. Section IV describes a method to achieve eye contact between a person and Robovie. Section V explains the utterance generation by Linta-III. Section VI describes a psychological experiment to evaluate the joint attention mechanism. Section VII discusses the results of the experiment and the effect of joint attention on human–robot interaction. Section VIII concludes the paper with a summary of our results and an overview of future work.

(a) (b)

Fig. 1. Joint attention and utterances.

II. Joint Attention in Dialogue

A. Joint Attention

People who are communicating with each other frequently focus on the same object. The focus of attention is called joint attention in social psychology [7]. This joint attention is a mental state where two people not only pay attention to the same information but also notice the other's attention to it. The following utterance example makes sense under joint attention.

Utterance Example 1:

R1-1: Please look at the poster.
R1-2: It's interesting.

In the example, a robot begins to explain a poster. Fig. 1 shows situations in which the above utterances are generated. In Fig. 1(a), the person is paying attention to only the robot, and not to the poster. In this scene, the robot mentions the poster with Utterance *R1-1*. In Fig. 1(b), the person comes to be aware of the poster and refers to it in his response to Utterance *R1-1*. The point of the example is that the person also becomes aware of the robot's attention to the poster. In short, the person and the robot come to have joint attention. According to the joint attention, the person catches the meaning of Utterance *R1-2* suggesting the poster.

The goal of this paper is to develop a speech generation system with joint attention. Since joint attention requires that a person and a robot focus on the same information, the speech generation system has to possess a mechanism to draw the person's attention to information focused on by the system.

B. Inference of Communicative Intention

People do not always have joint attention when they communicate with unfamiliar people. The restrictions on joint attention raise questions about whether people can actually have joint attention with robots, because robots have even fewer communicative relations than unfamiliar people.

Ono conducted a psychological experiment to investigate such questions [9]. The results of the experiment indicated that the relationship between a subject and a robot affects the joint attention development between them. For instance, in the experiment, a robot suddenly appeared in front of a subject and asked the subject to move a trash can blocking its way. The

subject was able to understand the request from the robot after being given a relationship with the robot. On the other hand, the subject was unable to understand the request description itself without the relationship being given. In short, the subject without a relationship did not have joint attention with the robot to the trash can because the subject ignored the unfamiliar robot and could not guess the robot's communicative intention, i.e. that it wanted to proceed forward. As a result, the speech generation system must possess a mechanism to develop a relationship between a person and a robot to achieve joint attention between them.

C. Attention in Generating Utterances

In this section, we discuss the attention representation used inside of a robot. Conventional attention mechanisms [3]–[5] are inadequate for generating situated utterances such as Utterance Example 1 because they are not able to grasp a person's attention. Instead, they simply deal with the focus of attention toward the nonverbal information (CG object or sensor information). For example, let's consider the following utterance example.

Utterance Example 2: Situation: in front of an obstacle

R2-1: I can't proceed forward.

Utterance *R2-1* does not require the person's attention to the obstacle because the utterance does not refer to any real-world object. If the robot has sensors to detect the obstacle, even a conventional attention mechanism can generate the above utterance. In contrast, the attention mechanism must grasp the person's attention to decide whether it is possible to generate the following utterance.

Utterance Example 3: Situation: in front of a person and an obstacle

R3-1: Please move this!

Utterance *R3-1* asks the person to move the obstacle, and the person's attention to the obstacle is required for him/her to notice that the word "this" refers to the obstacle. This example suggests that the person's attention is vital for the speech generation system to generate a situated utterance description. As a result, the attention mechanism must deal with the person's attention.

D. Robovie

To develop a speech generation system, we employed the humanoid robot named Robovie [Fig. 2(a)], which we had previously developed. Robovie has two four–degrees–of–freedom (4DOF) hands and a 3DOF head to generate hand gestures and gaze with head motions. It also has a movable base and several types of sensors: touch sensors to notice being touched by persons, ultrasonic distance sensors to detect obstacles, and two types of vision sensors. One of the vision sensors is an omnidirectional vision sensor that can capture panoramic scenes around Robovie. Robovie obtains the direction of a person's location by detecting skin color movements in an image captured by the omnidirectional vision sensor. The other vision sensor uses two charge-coupled device (CCD) cameras on Robovie's head for stereo vision. Robovie is also equipped with speech recognition and generation software.

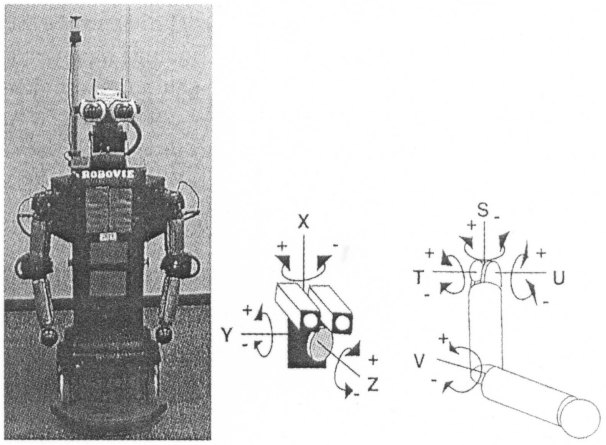


Fig. 2. Everyday robot "Robovie."

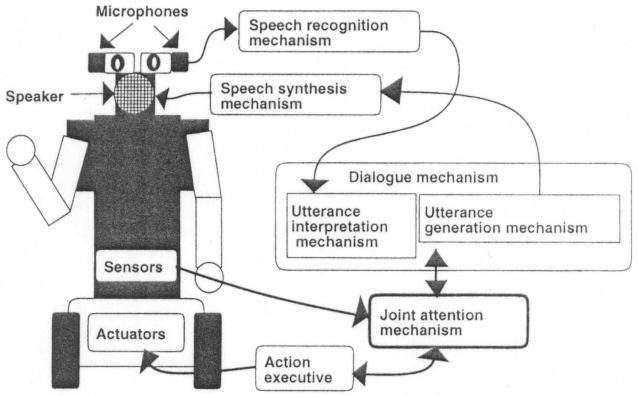


Fig. 3. Overview of Linta-III.

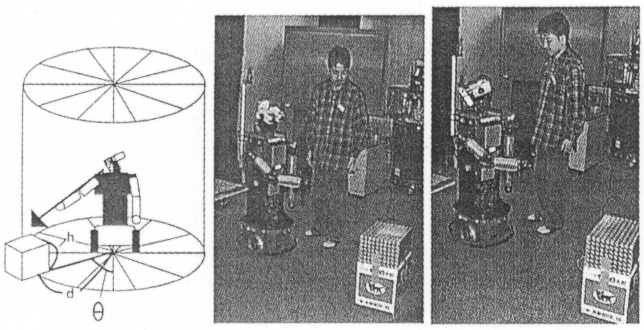
To explain the gestures of Robovie in the remaining sections, we define representations of Robovie's head position and hand position. The head position is defined as $Head(X, Y, Z)$, where X, Y, and Z denote axes to turn the head to the left or right, to move the head up or down, and to put the head to one side, respectively [Fig. 2(b)]. The hand position is defined as $Hand_i(S, T, U, V)$, where i indicates the left $(i = l)$ or right $(i = r)$ hand. Fig. 2(b) shows the movable point of each axis, S, T, U, and V.

III. Linta-III and Joint Attention Mechanism

In this paper, we propose a speech generation system named Linta-III and a joint attention mechanism to develop joint attention between a person and Robovie.

Fig. 3 shows an overview of Linta-III. Linta-III consists of the joint attention mechanism, a dialogue mechanism (an utterance generation mechanism and an utterance interpretation mechanism), an action executive, Robovie's actuators, and Robovie's sensors. In Linta-III, the joint attention mechanism develops joint attention between a person and Robovie by using Robovie's sensors and actuators. According to the developed joint attention, Linta-III can generate a situated utterance description with the utterance generation mechanism.

The joint attention mechanism manifests Robovie's focus of attention with an attention expression function to draw the



(a) (b) (c)

Fig. 4. Attention expression and eye contact by Robovie.

person's attention. The attention expression function indicates source of the sensor information under focus by using gaze motions and hand gestures [Fig. 4(b)].

The joint attention mechanism also employs an eye-contact function to develop a communicative relationship between the person and Robovie and to force the person to guess the communicative intention of the attention expression. From an awareness of the attention expression, the person comes to focus on the same real-world information as Robovie. Eye contact is achieved by turning Robovie's head to face the person [Fig. 4(c)].

Moreover, the joint attention mechanism possesses an attention coordinate. The attention coordinate grasps the sensor information focused on not only by Robovie but also by the person. In other words, the attention coordinate deals with the joint attention between the person and Robovie. According to the current attention coordinate, Linta-III decides whether the utterance description depends on a situation.

A. Attention Expression

The attention expression function works by pointing out the source of the sensor information under focus by using head and hand motions. The attention expression mechanism represents the sensor information on a cylindrical coordinate (θ, d, h) (attention coordinate). All of Robovie's sensor information is put on the coordinate to easily carry out the attention expression.

Fig. 4(a) shows an example of the attention expression. In the figure, Robovie manifests its attention to the box, which is located at a distance d and a direction θ from Robovie. The height of the box is h. According to the location of the box, the pose of the head becomes $Head(\theta, \arctan(d/(L-h)), 0)$ to face the box. Here, L is the height of Robovie's head. In addition, the pose of the right hand becomes $Hand_r(0, \arctan(\cos\theta \tan\theta_1), \arcsin(\sin\theta_1 \sin\theta), 0)$ to point out the box. Here, θ_1 is the angle $\arctan(d/(H-h))$ between the hand and the body of Robovie. H is the height of Robovie's shoulder.

B. Relation With a Person Based on Eye Contact

The eye-contact function turns Robovie's gaze to the person's direction. The person's location detected with the omnidirectional vision sensor is represented as (θ_p, _, T) in the attention coordinate. θ_p gives the person's direction from Robovie's front, and T is the height of the omnidirectional vision sensor from ground level. According to the person's location, the position of the head becomes $Head(\theta_p, \arctan(D/T), 0)$ to achieve eye contact. Here, D denotes the maximum distance within which the omnidirectional vision sensor can obtain the person's location.

The joint attention mechanism frequently carries out eye contact during the attention expression to make the person notice the expression. For example, when Linta-III generates Utterance Example 1, it turns Robovie's head to the poster and then to the person while making Robovie's hand point to the poster. As a result of the eye contact, the person guesses the communicative intention of Robovie and becomes aware of Robovie's attention.

C. Attention Toward Sensor Information

We add information f to the attention coordinate to express the type of object in the sensor information. The following representation gives an improved attention coordinate:

$$f\,[(\theta, d, h)]\,. \tag{1}$$

There are four object types in Linta-III: a person p, an object o, a poster e, and Robovie r. For example, an obstacle located at a distance of 50 cm and an angle of $-30°$ from Robovie's front is expressed as $o[(-30, 50, U)]$ in the attention coordinate. Here, U is the height of the ultrasonic distance sensors from ground level.

The attention coordinate for a person p has the following different expression to represent the person's attention:

$$p\,[(\theta_p, \_, T), a]\,. \tag{2}$$

Here, a is a slot for the person's attention. For example, if the person is paying attention to the obstacle in the previous paragraph, the attention coordinate becomes $p[(\theta_p, \_, T), (o[(-30, 50, U)])]$.

The joint attention mechanism alters the attention coordinate depending on the attention expression and eye-contact functions. We now explain the alteration for Utterance Example 1. In this example, Robovie pays attention to direction θ_p of person p and direction θ_e of poster e. According to the attention, the joint attention mechanism has the following two attention coordinates:

$$p\,[(\theta_p, \_, T), \_]\,, \quad e\,[(\theta_e, \_, T)]\,. \tag{3}$$

Direction θ_e of the poster is also obtained from the omnidirectional vision sensor by color detection.

To explain the poster, the joint attention mechanism turns Robovie's head to the person and carries out eye contact. Since the person notices the existence of Robovie through the eye contact, the joint attention mechanism adds the existence of Robovie $r[(0, 0, 0)]$ to the person's attention coordinate and obtains the following coordinate:

$$p\,[(\theta_p, \_, T), (r\,[(0, 0, 0)])]\,, \quad e\,[(\theta_e, \_, T)]\,. \tag{4}$$

Here, the reason why each value of Robovie r becomes 0 is that the location of Robovie itself is 0.

Next, the joint attention mechanism manifests Robovie's attention to the poster by turning Robovie's gaze to it and indicating it with Robovie's hand. Since the person is made to pay attention to the poster by the attention expression, the joint attention mechanism adds $e[(\theta_e, \_, T)]$ to the person's attention coordinate

$$p\left[(\theta_p, \_, T), (r\left[(0,0,0)\right],\ e\left[(\theta_e, \_, T)\right])\right], \quad e\left[(\theta_e, \_, T)\right]. \tag{5}$$

As a result, the joint attention mechanism achieves joint attention to the poster $p[(\theta_p, \_, T), (r[(0,0,0)], e[(\theta_e, \_, T)])]$.

IV. Eye Contact With a Person

The joint attention mechanism achieves eye contact by recognizing a person's location and turning Robovie's head in that direction. The location recognition function consists of two recognition stages: a recognition stage with the omnidirectional vision sensor and a recognition stage with the ultrasonic distance sensors.

The omnidirectional vision sensor employs motion information calculated from interframe differences and skin color information to track the person's location. This method has the advantages of computing at lower costs and recognizing the person more robustly against changes in illumination. The omnidirectional vision sensor recognizes the person's location in the following two steps. First, it seeks out a different region by comparing two omnidirectional images in the time sequence. Next, it compares color information (RGB values) based on skin color at the region found to be different. If there is such a region found by the above two steps, the omnidirectional vision sensor calculates the direction θ_p of the located person. Since the shape of the omnidirectional image is concentric circles, it is easy to obtain the direction information from the found region. However, the person's location recognition is performed only when Robovie does not wander. The reason for this is that, since all regions of the omni-directional image moves when Robovie wanders, the omidirectional vision sensor cannot find interframe differences appearing in the part of the image.

After the omidirectional vision sensor finds the person, the joint attention mechanism also examines the person's location with the ultrasonic distance sensors. In short, it recognizes the person's existence by using the distance information. Robovie has 16 ultrasonic distance sensors around the movable base (the lower part of Robovie) and eight ultrasonic distance sensors around its torso (the upper part of Robovie). The sensor values are dealt with based on the following two points: whether the upper part and lower part sensors in the direction θ_p have similar distance values and whether the sensor values in the direction θ_p are smaller (which means that someone/something is located near Robovie) than the values of the next sensors. As a result, the joint attention mechanism recognizes the person's existence with both the vision sensor and the distance sensors.

In addition, the design of a robot's head and behavior is also vital to achieving eye contact. The two cameras on Robovie's head and the speed of the head movement is designed to draw a person's attention to Robovie's gaze (180°/s).

TABLE I
Utterance Generation Rules

Rule 1	$v(\_, f)$	$\rightarrow$	$v(r, \_)$
Rule 2	$v(\_, f), p[\_, \_]$	$\rightarrow$	$Hello, v(\_, f)$
Rule 3	$v(\_, f), p[\_, (r, \_)]$	$\rightarrow$	$v(\_, f)$
Rule 4	$v(\_, f), p[\_, (r, f)]$	$\rightarrow$	$v(\_, dp)/\ v(\_, \_)$

V. Interaction With Robovie

A. Utterance Generation Rules

Linta-III generates an utterance by referring to the attention coordinate. Since the utterance generation and the joint attention mechanism are carried out concurrently, Linta-III is able to generate an utterance independent of the control of eye contact and attention expression. It generates the situated utterance in response to the current attention coordinate.

Linta-III has the utterance generation rules listed in Table I. To generate a situated utterance description, Linta-III selects an appropriate rule for the current attention coordinate. The rules in Table I take the form of (*utterance content*), (*attention coordinate*)$\rightarrow$ (*altered utterance content*). The utterance content takes the form of $v(subject, object/complement)$. For example, the utterance "please look at the poster!" takes the form of $look(p, e)$. Since the target of Linta-III is joint attention, only attention to a person $p[\_, \_]$ is listed in Table I. r represents the existence of Robovie, f any type of object, and dp a demonstrative pronoun. In addition, $\_$ in the rules is able to take any value. The rules also omit location information from the attention coordinate.

Linta-III uses Rule 1 when no person is within range of the omnidirectional vision sensor in generating an utterance. Because there is no person around Robovie at the moment, the rule generates the utterance $v(r, \_)$, which describes the situation from Robovie's viewpoint to report it to any person passing there by chance. Utterance Example 2 corresponds to the use of Rule 1. In the example, since there is no person p to ask to move obstacle o, Robovie r reports $\neg proceed(r)$, which means that it cannot proceed forward. The example can be represented as $move(p, o) \rightarrow \neg proceed(r)$ by Rule 1.

Linta-III uses Rule 2 when a person is within range of the omni-directional vision sensor but has no communicative relationship with Robovie. At first, Rule 2 has Robovie greet the person to develop a relationship. After the greeting, Rule 2 generates an utterance without omitting sensor information f because the person is most likely not paying attention to f in the situation.

Linta-III uses Rule 3 for a person who has already established a relationship with Robovie by eye contact. However, since the person is most likely not paying attention to f (as in Rule 2), Rule 3 does not omit f from the utterance description. Utterance Example *R1-1* corresponds to a use of Rule 3. In the example, Linta-III tries to ask person p to look at poster e. However, since the person is not paying attention to the poster, Linta-III does not omit the word "poster (e)" from the utterance content $look(p, e)$. The example can be represented as $look(p, e)$, $p[\_, (r)] \rightarrow look(p, e)$ by Rule 3.

Linta-III uses Rule 4 after joint attention has been achieved by the attention expression and eye-contact functions. Under

the joint attention, Linta-III uses a demonstrative pronoun instead of f (or sometimes completely omits it from an utterance description). Utterance Examples *R1-2* and *R3-1* correspond to uses of Rule 4. In *R1-2*, Linta-III tries to inform person p that poster e is interesting. Since joint attention to the poster $p[\_, (r, e)]$ has already been achieved in the example, Linta-III omits e from the utterance content $is - a(e, interesting)$ with the demonstrative pronoun "it." The altered utterance content is $is - a(it, interesting)$. The example can be represented as $is - a(e, interesting)$, $p[\_, (r, e)] \rightarrow is - a(it, interesting)$ by Rule 4.

B. Utterance Generation With Joint Attention

In this section, we demonstrate several patterns of utterance generation for Utterance Example 3, each of which is adapted to fit the attention coordinate by rule selection.

In the situation of Utterance Example 3, Robovie is blocked by an obstacle. At that time, Linta-III uses utterance content $move(p, o)$ to ask a person to move the obstacle. If the omnidirectional vision sensor does not capture the existence of a person, the attention coordinate comes to have only $o[\_]$, which denotes the existence of an object. Since the attention coordinate does not have $p[\_, \_]$, Linta-III selects Rule 1 and has utterance content $\neg proceed(r)$. As a result, Linta-III generates Utterance Example 2.

If the omnidirectional vision sensor captures a person's location $(\theta_p, \_, T)$, the attention coordinate becomes $p[(\theta_p, \_, T), \_]$. Then, the joint attention mechanism moves Robovie's head to carry out eye contact. However, eye contact is not always achieved, particularly if the person is moving quickly. If Linta-III begins to generate an utterance before achieving eye contact, it selects Rule 2 $move(p, o)$, $p[(\theta_p, \_, T), \_] \rightarrow Hello$, $move(p, o)$. As a result, Linta-III generates the utterance "Hello! Please move the obstacle."

If eye contact is achieved, the attention coordinate becomes $p[(\theta_p, \_, T), (r(0, 0, 0), \_)]$. Then, Linta-III selects Rule 3 $move(p, o)$, $p[(\theta_p, \_, T), (r(0, 0, 0), \_)] \rightarrow move(p, o)$. As a result, it generates only one utterance "Please move the obstacle."

If eye contact and attention expression are achieved before an utterance is generated, the person is assumed to be paying attention to the object. Then, the attention coordinate becomes $p[(\theta_p, \_, T), (r(0, 0, 0), o(\theta_o, d, h))]$. In response to the attention coordinate, Linta-III selects Rule 4 $move(p, o)$, $p[(\theta_p, \_, T), (r(0, 0, 0), o(\theta_o, d, h))] \rightarrow move(p, this)$. As a result, it generates Utterance Example 3. Here, the demonstrative pronoun "this" is selected based on the distance between Robovie and the obstacle [10].

VI. Psychological Experiment

A psychological experiment was conducted to verify the effect of eye contact on achieving joint attention. The experiment divided 20 subjects into two equal groups; one was given Robovie with eye contact and the other was given Robovie without eye contact in interaction. The joint attention mechanism carried out attention expression for both groups. The target of the attention expression was a poster on a wall. The experiment recorded the number of subjects who looked at the poster in response to the attention expression.

TABLE II
Experimental Results: Comparison of the Number of People Who Looked at the Poster Indicated by Robovie According to the Effect of Eye Contact ($U = 5, p < .01$)

	Looked at poster	Looked at Robovie's hand
With eye-contact	10	0
Without eye-contact	1	9

The experimental procedure was carried out as follows. First, Robovie passed in front of the subject and stopped in front of the poster, where both Robovie and the poster were in the subject's sight. At the location, Robovie turned to the subjects and pointed to the poster with its hand while generating the utterance "please look at this." At this point, for half of the subjects, Robovie repeatedly carried out eye contact with the subjects and attention expression to the poster. For the remaining half of the subjects, Robovie did not carry out eye contact but instead had its head fastened in the forward direction.

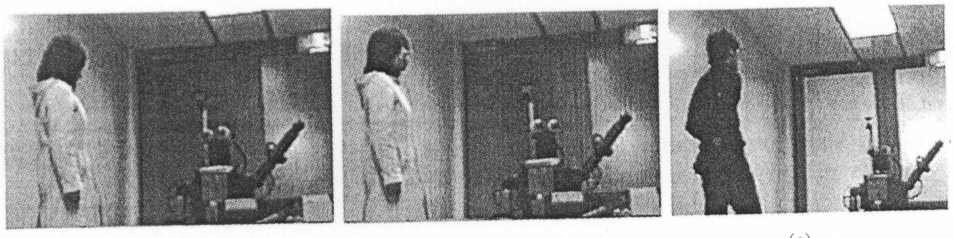
Table II shows the results, which indicate that all subjects with eye contact [Fig. 5(a)] looked at the poster [Fig. 5(b)]. On the other hand, all but one of the subjects without eye contact looked at Robovie's hand instead of the poster [Fig. 5(c)]. As a result, there was significance in the effect of eye contact on achieving joint attention ($U = 5, p < .01$).

VII. Discussion

The experimental results indicate that the subjects without eye-contact did not become aware of Robovie's ability to refer to a third object in the environment. In short, the relationship between the subjects and Robovie was binomial, i.e., there was no relationship with items in the environment. In the past, communication between people and conventional robots was based on such a binomial relationship. On the other hand, with eye contact, the subjects become aware of Robovie's ability to point out an object and accordingly looked at the poster. This finding of reference to the poster indicates that a triadic relationship involving the subjects, Robovie, and the poster emerged instead of a binomial relationship. In short, Linta-III achieved a new type of communication between people and a robot based on a triadic relationship.

In addition, the joint attention mechanism communicates Robovie's attention to a person through both a physical relation by eye contact and an attention expression by physical behavior. On the other hand, the relevance theory [6] explains communications only based on reasoning. ¿From the experimental results, it is obvious that a relationship developed by eye contact has a more fundamental effect on communications than logical reasoning or knowledge processing. Since Linta-III is able to communicate with people by referring to real-world information with physical behaviors, it has the potential to achieve more natural human–robot interaction.

The joint attention mechanism employs the eye-contact function to achieve joint attention. Actually, several other studies have employed eye contact for human–robot interaction



(a) (b) (c)

Fig. 5. Experimental scenes: (a) eye contact between a person and Robovie, (b) joint attention to the poster, and (c) a person looking at Robovie's hand (in right photo).

[11]–[13]. In the following three paragraphs, we explain the differences between how we exploit eye contact and how it was used in the previous research efforts.

Humanoids named Kizmet [2] and Cog [12] identify the target of the person's attention by recognizing the location of a person's face and the orientation of the face through vision processing. In these humanoids, eye contact is employed as a social cue to identify the attention target because a person carries out eye contact before turning his gaze to the target. Moreover, a humanoid named Infanoid [11] recognizes the person's gaze direction by extracting eye areas from a visual image. To achieve joint attention, the above three humanoids adjust their attention to the target focused on by the person. In short, the joint attention may be regarded as human-centered joint attention. However, when a robot plays the lead role in a communication (e.g., an information robot like a route guide robot), another approach to joint attention is necessary for human–robot interaction. This is robot-centered joint attention.

The originality of Linta-III is its pursuit of robot-centered joint attention. The big difference from the conventional research approaches is that eye contact is employed in the joint attention mechanism as a social representation toward a person. As a result of such a representation, the person understands the target focused on by Robovie and pays attention to it as the experimental results suggest.

Moreover, the joint attention mechanism employs eye contact as the social representation to develop a triadic relationship between the person, Robovie, and a physical object (i.e., the poster in the example). Although another study also employed eye contact as social representation [13], the interaction dealt was based on a binomial relationship. This robot expresses somebody's turn to talk with it by carrying out eye contact with one of plural persons. In addition, a humanoid robot named SIG [14] integrates audio information (voice direction) with visual information (face direction) to track the speaking person by turning its head. The communication between the person and SIG is also not based on triadic relationship. In contrast to these research projects, our joint attention mechanism employs eye-contact expression to develop a triadic relationship.

The joint attention mechanism carries out eye-contact by turning Robovie's head to the person's location recognized with the omnidirectional vision sensor and the ultrasonic distance sensors. On the other hand, the joint attention mechanism does not confirm the achievement of eye contact in terms of the gaze direction of the person. As the experimental results suggest, the achievement of joint attention goes well when Robovie turns its head to both the person's direction and the target of attention. However, if the joint attention mechanism recognized the person's gaze, it would be able to control the time when Robovie carries out attention expression. To improve the achievement of joint attention, a recognition method of the person's gaze must be developed in the future.

VIII. Conclusions

This paper proposed a speech generation system called Linta-III installed on a humanoid robot named Robovie. Linta-III has a joint attention mechanism to refer to real-world situations in generating an utterance. Since the joint attention mechanism can draw a person's attention to the same real-world information as focused on by Robovie, Linta-III can generate a situated utterance description by omitting obvious words in a situation.

The joint attention mechanism employs eye-contact between the person and Robovie, and an attention expression (e.g., pointing to sensor information by hand gestures and using Robovie's gaze to draw a person's attention). It also has an attention coordinate to represent joint attention. By referring to the attention coordinate, Linta-III can adapt an utterance description to suit the process of developing the joint attention. In addition, we conducted a psychological experiment. The experimental results clearly indicated that the effect of eye contact on achieving joint attention has significance.

In the future, we will employ the joint attention mechanism in recognizing a situated utterance from a person. Moreover, we must study how to develop joint attention to real-world information that a person is focusing on by detecting the person's gaze or inferring it from the interaction context.

References

[1] J. Barwise and J. Perry, *Situations and Attitudes*. Cambridge, MA: MIT Press, 1983.
[2] C. Breazeal and B. Scasselati, "A context-dependent attention system for a social robot," in *Proc. IJCAI'99*, 1999, pp. 1146–1151.
[3] M. Imai, K. Hiraki, and Y. Anzai, "Human-robot interface with attention," *Tranc. Inst. Electron. Inf. Commun. Eng. D*, vol. J77-D-II, pp. 1447–1456, 1994.
[4] D. Chapman, *Vision, Instruction, and Action*. Cambridge, MA: MIT Press, 1991.
[5] B. J. Grosz, "The representation and use of focus in a system for understanding dialogs," in *Proc. IJCAI'77*, 1977, pp. 67–76.
[6] D. Sperber and D. Wilson, *Relevance: Communication and Cognition*. Oxford, U.K.: Blackwell, 1986.
[7] C. Moore and P. J. Dunham, *Joint Attention: Its Origins and Role in Development*. Hillsdale, NJ: Lawrence Erlbaum, 1985.

[8] B. Scassellati, "Imitation and mechanisms of joint attention: a developmental structure for building social skills on a humanoid robot," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer, 1999, vol. 1562, pp. 176–195.
[9] T. Ono and M. Imai, "Reading a robot's mind: A model of utterance understanding based on the theory of mind mechanism," in *Proc. AAAI 2000*, 2000, pp. 142–148.
[10] M. Imai, K. Hiraki, and T. Miyasato, "Physical constraints on human robot interaction," in *Proc. IJCAI'99*, 1999, pp. 1124–1130.
[11] H. Kozima and A. Ito, "Toward mindreading by an attention-sharing robot," in *Proc. Third Int. Symp. Artificial Life and Robotics (AROBO III'98)*, 1998, pp. 478–481.
[12] R. A. Brooks, C. Breazeal, M. Marjanović, B. Scassellati, and M. M. Williamson, "The Cog project: Building a humanoid robot," in *Computation for Metaphors, Analogy and Agents*, C. Nehaniv, Ed. Berlin, Germany: Springer, 1998, vol. 1562, Springer Lecture Notes in Artificial Intelligence.
[13] Y. Hidaki, K. Masumitsu, N. Yamagishi, Y. Nakano, N. Kobayashi, S. Haruyama, T. Kobayashi, and A. Takanishi, "A robot who converse with plural persons using eye-contact," in *Proc. IPSJ GIG-Notes Spoken Language Processing*, vol. 97-SLP-17-1, 1997, pp. 1–6.
[14] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano, "Real-time auditory and visual multiple-object tracking for humanoids," in *Proc. IJCAI 2001*, 2001, pp. 1425–1432.

Michita Imai received the Ph.D. degree in computer science from Keio University, Yokohama, Japan, in 2002.

He is a Research Associate with the Faculty of Science and Technology at Keio University and a Researcher with the ATR Media Information Science Laboratories, Kyoto, Japan. In 1994, he joined the NTT Human Interface Laboratories. He joined the ATR Media Integration and Communications Research Laboratories in 1997. His research interests include autonomous robots, human–robot interaction, speech dialogue systems, humanoids, and spontaneous behaviors.

Dr. Imai is a member of the Institute of Electrical, Information and Communication Engineers, Japan, Information Processing Society of Japan, Japanese Cognitive Science Society, Japanese Society for Artificial Intelligence, Human Interface Society, and Association for Computing Machinery.

Tetsuo Ono received the Ph.D. degree in information science from the Japan Advanced Institute of Science and Technology, Fukui, Japan, in 1997.

He is currently an Associate Professor of Media Architecture at the Future University, Hakodate, Japan. He was a Researcher at the ATR Media Integration and Communications Research Laboratories between 1997–2001. His research interests include computational mechanisms of emotion, evolution of languages, human–robot communications, and interactive systems.

Dr. Ono is a member of the Information Processing Society of Japan, Japanese Cognitive Science Society, and Japanese Society for Artificial Intelligence.

Hiroshi Ishiguro received the D.Eng. degree from Osaka University, Osaka, Japan, in 1991.

In 1991, he was a Research Assistant in the Department of Electrical Engineering and Computer Science, Yamanashi University, Japan. He then joined the Department of Systems Engineering, Osaka University, as a Research Assistant in 1992. In 1994, he was an Associate Professor in the Department of Information Science, Kyoto University, Japan, and began research on distributed vision using omnidirectional cameras. From 1998 to 1999, he was with the Department of Electrical and Computer Engineering, University of California, San Diego, as a Visiting Scholar. In 1999, he was a Visiting Researcher at the ATR Media Information Science Laboratories, where he helped develop the interactive humanoid robot, Robovie. In 2000, he moved to the Department of Computer and Communication Sciences, Wakayama University, Japan, as an Associate Professor. He became a Professor in 2001. He was also a Researcher with PREST, Japan Science and Technology Corporation in 2001. He is currently a Professor in the Department of Adaptive Machine Systems, Osaka University.