Human-Like Conversation with Gestures and Verbal Cues based on Three-Layer Attention-drawing Model

Osamu Sugiyama, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro and Norihiro Hagita
ATR Intelligent Robotics and Communication Laboratories, 2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, JAPAN
Email: sugiyama@atr.jp, kanda@atr.jp, michita@ayu.ics.keio.ac.jp.

Abstract:

When we talk about objects in an environment, we indicate to a listener which object is currently under consideration by using pointing gestures and reference terms as 'this' and 'that'. Such reference terms play an important role in human interaction by quickly informing the listener of an indicated object's location. In this research, we propose a three-layer attention-drawing model for humanoid robots that incorporates gestures and verbal cues. Our proposed three-layer model consists of three sub-models: the Reference Term Model (RTM), the Limit Distance Model (LDM), and the Object Property Model (OPM). The RTM selects an appropriate reference term using functions constructed by an analysis of human behavior. The LDM decides whether to use another property of the object, such as color, with a reference term for identifying the object. The OPM determines the most appropriate property for indicating the object by comparing object properties with each other. We have developed an attention-drawing system in a communication robot named 'Robovie' based on this model. We confirmed its effectiveness through experiments.

Keywords: Human Robot Interface; Human Robot Interaction; Deictic Gestures;

1 INTRODUCTION

Recently, many humanoid robots have been developed (Trafton et al. 2005, Kidd et al. 2004, Kamasima et al. 2004, Nakadai et al. 2001) and in addition to these, there is work underway to develop android robots that have human-like appearance (Walters el al. 2005, Kanda et al. 2005, Goetz et al. 2003). Such development is leading to the emergence of research into 'android science' (Ishiguro 2005). Android science is based on the premise that an android's human-like appearance would enable us to engage in natural communication with it. At the same time, we can understand more about 'what is a human?' by implementing such a communication capability into a human-like robot. There are several research works conducted on human-like motion, such as blinking, mouth movement when it is speaking, and the combination of natural arm and head movements (Sakamoto et al. 2005). Moreover, we believe that it is also important to make robots capable of talking about objects in a daily environment with a human-like casual manner.

Several previous studies have demonstrated that humanoid robots are capable of human-like conversation about certain objects. For example, Scassellati et al. have developed a humanoid robot that performs pointing and gaze-following motion (Scassellati 2000, Breazeal et al. 2000), which is well known as 'joint-attention' (Moore et al. 1995). Imai et al. also implemented the joint-attention mechanism on a humanoid robot and proved its effectiveness experimentally (Imai, 2003). The robot draws the listener's attention by establishing

eye contact with the listener, then looks at the target object to talk about it, pointing at the object while saying 'look at this'. As the examples in these works show, a robot can engage in human-like conversation about objects by effectively using its human-like body. They demonstrated the effectiveness of using their humanoid body in communication.

However, in these studies, the positions of the objects were static and pre-programmed into the robot, meaning these robots were not capable of conversation about different objects or in different environments. One of the difficulties in this field is choosing appropriate reference terms along with the positions of objects, such as 'this' or 'that' (in Japanese, there are three types of reference terms and the usage is more complex; we will explain this in Section 2). Another difficulty is determining when it is appropriate to use a simple expression, such as 'pick this up (with a pointing gesture)' and when to use more complex expressions, such as 'pick up the large black box near the wall to your left'.

Our approach is based on a three-layer model. The first layer deals with the relationship between reference terms and positions of objects. In Japanese, this relationship is complicated and is determined according to the position of the speaker, the listener and the object about which they are talking. The second layer addresses the robot's pointing gestures, which depend on the robot's physical form. The third is about properties of objects in order to identify each object with language. The first two layers are used to determine whether the robot can talk about an object with simple expressions and the third layer is mainly used when it cannot use such simple expressions. These three layers were prepared based on observations of human behavior. The effectiveness of the system based on the three-layer model will be demonstrated in experiments.

2 . RELATED WORKS

In our research, we deal with the behavior that gestures and speech complement each other. In "Psycholinguistics – A new approach," (McNeill, 1990), there is a sentence stating that, "It is necessary for demonstrative pronouns such as 'this' and 'that' to be used with gestures in order to make those words grammatically precise." (Levelt et al., 1985) A pointing gesture limits the space in which the object exists based on the pointing direction. On the other hand, a reference term limits the space in which the object exists according to the positions of the speaker and listener. By using these different types of spatial limitation, the behavior can draw the listener's attention toward the indicated object more accurately. We have named this behavior "Attention Drawing Behavior" and aim to further develop human-robot interaction using this behavior.

There are many related studies on attention-drawing interaction in robotics. One type of research focuses on joint-attention in relation to child development (Kojima & Nagai, 2005); also, there are studies on mutual non-verbal behavior, such as research toward human-robot teamwork (Breazeal, 2005). However, these investigations do not address a method of using reference terms appropriately in various situations.

Several previous studies focused on a method for robots to recognize human pointing gestures or utterances with reference terms. For instance, Mizuno et al. proposed a method that informs robots of an object's location both hand gestures and verbal cues (Mizuno, 2003), while Haasch et al. proposed a method to recognize an object by using a pointing gesture, a human utterance, and stored information on objects (Haasch, 2005). Meanwhile, Hanafiah at el. made a robot recognize an object using inexplicit utterances including reference terms and pointing gestures (Hanafiah, 2004). Other studies have focused on a recognition method that uses utterance contents such as reference terms and information on the object in question. For instance, Inamura et al. proposed a probabilistic method of recognizing an object indicated by a human, using object

information such as color or size in utterances with combination with reference terms (Inamura, 2004). In these studies, recognition is, however, investigated separately from the attention-drawing mechanism of a robot.

On the other hand, while many robots are equipped with an attention-drawing mechanism (Scasselati, 2000; Imai, 2003), they cannot dynamically handle environments where the locations of objects and people change; that is, they only performed pre-implemented gestures and utterances.
To tackle these problems we propose a three-layer attention-drawing model for a robot that incorporates pointing gestures and reference terms.

3. ANALYSIS OF HUMAN BEHAVIOR IN REFERRING TO OBJECTS

We observed inter-human conversation that refers to objects in a certain environment in order to develop a human-like conversation capability for a robot. This chapter describes the result of three analyses of human conversation. The first is an analysis of the effect of positional relationships among speakers, listeners, and objects regarding the use of reference terms. The second analysis focuses on the situation where two or more objects become close together, and the third is an analysis of the situation where humans cannot identify the target object by a reference term only.

3.1 Conversation using reference term

3.1.1 Reference Terms in Japanese

In Japanese, we use three reference terms to inform people of an object's location: *kore*, *sore*, and *are*. They correspond to 'this' and 'that' in English, with *kore* referring to an object close to the speaker, and *sore* to an object close to the listener or between the two individuals. *Are* refers to an object that is neither close to the speaker nor to the listener. However, these definitions are so vague that for the sake of consistency, we have to set up some borders to classify the usage of these reference terms.

3.1.2 Investigating the usage of reference terms

To investigate the spatial factors that determine which reference term should be used, we conducted an experiment to observe inter-human conversations referring to an object in the environment, as shown in Fig. 1.

Fig.1: Analysis of Human Behavior

3.1.2.1 Experimental Procedure

[Brief Overview] We set up an environment in which a speaker, a listener, and an object exist. In the experiment, the speaker asked the listener to pick up a round object (11cm radius and 30cm tall) using one of the following expressions in Japanese:

- '*kore-totte*' ('pick up this')
- '*sore-totte*' ('pick up that')
- '*are-totte*' ('pick up that')

We changed the object's location and recorded which reference term the speaker used in each trial.

[Subjects] The subjects were 20 male & female university students. They were divided into 10 pairs: one played the speaker role and the other the listener role.

[Experiment Procedure] The experiments consisted of seven sessions. In each session, we changed the distance between speaker and listener and their direction. The distance had three variations: 50 cm, 1.0 m, and 2.0 m, and there were four variations of direction between the speaker and the listener as shown in Fig. 2. The combination of distances and angles are as follows:

1. The distance is 1.0 m and the angle is pattern (a) in Fig. 2.
2. The distance is 1.0 m and the angle is pattern (b) in Fig. 2.
3. The distance is 1.0 m and the angle is pattern (c) in Fig. 2.
4. The distance is 1.0 m and the angle is pattern (d) in Fig. 2.
5. The distance is 0.5 m and the angle is pattern (c) in Fig. 2.
6. The distance is 0.5 m and the angle is pattern (d) in Fig. 2.
7. The distance is 2.0 m and the angle is pattern (a) in Fig. 2.

The subjects participated in all sessions in the experiment and the order of the sessions differed from one individual to another. In each session, the speaker asked the listener to pick up the object using one of the utterances described above. There were some rules in the trials: first the speakers could twist their waist and neck in order to refer to the object; second, they could use a pointing gesture if necessary to refer to the object. After the speaker had referred to the object, the experimenter moved the object to another position and repeated these steps in all trials of the session.

[Analysis Method] We calculated the probability of each reference term used for each object's location. Based on these probabilities, for each angle we calculated the distance where the probabilities of both reference terms became 50%, and defined it as a border for determining the usage of each reference term.

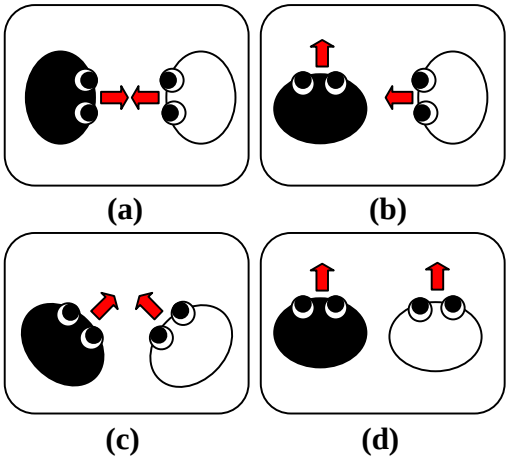
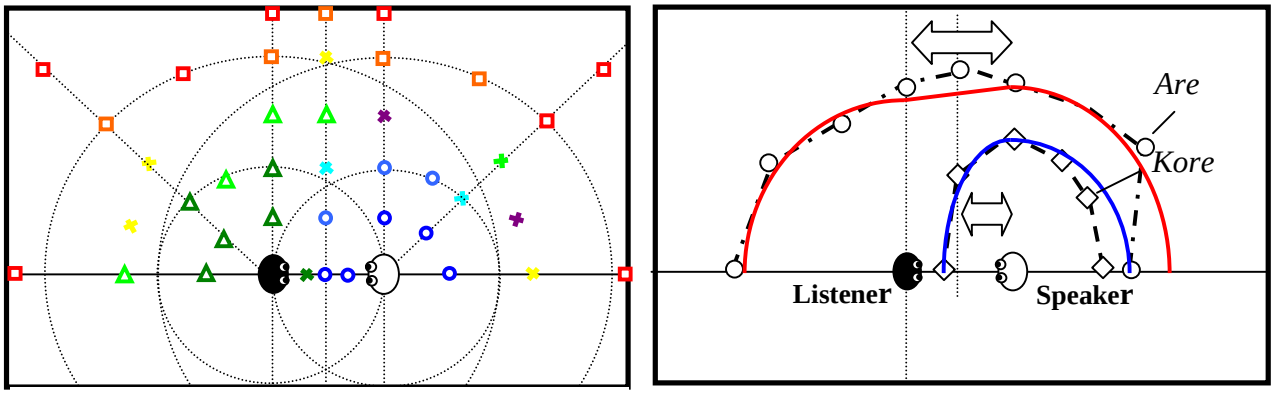


Fig.2: The directions of subjects

3.1.2.2 Experimental Result

Figure 3 shows the positions of the object in Session 1 and a border for determining the usage of each reference term.



(a) the object positions in Session 1 (b) the calculated border in Session 1

Fig.3: the actual object positions in Session 1 and its calculated border

In Fig. 3 (a), the circles mean that the reference term *kore* is most frequently used in that position, the triangles mean that *sore* is most frequently used, and the squares denote that *are* is most frequently used. The positions marked by crosses indicate that all reference terms are used with equal frequency. On the other hand, in Fig. 3 (b) we illustrate the points where the probability of using *kore* and *are*, or *sore* and *are*, becomes 50%. Each point is connected by a line, and we also illustrate the approximate curve for the reference term border calculated using the least-squares method.

Based on the results in each Session, in Fig. 4 we produced a rough sketch of the borders between the regions where each reference term is used: the *kore* region, the *sore* region, and the *are* region. For example, when an object is in the *kore* region, the speaker mainly uses the reference term *kore* to refer to the object. The border between *kore* and *sore* is an ellipse with its center at the speaker, and the border between *sore* and *are* is an ellipse whose center is between the speaker and the listener.

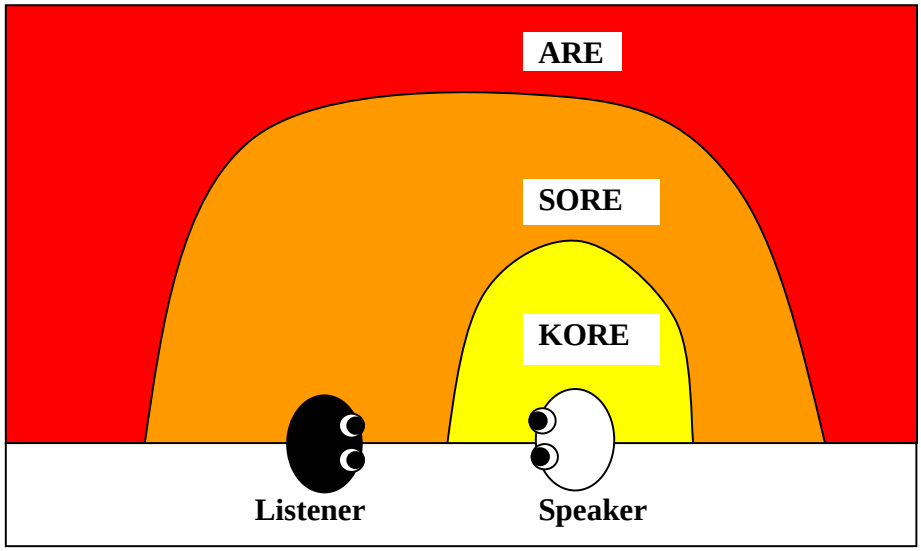


Fig.4: The usage region of each Japanese Reference Term

3.2 Conversation when two or more objects are close together

When two or more objects in the environment are close together, it is difficult for the listener to identify the object only by using a pointing gesture and a reference term. We define Limit Distance as the distance at which we cannot distinguish the indicated object from the other objects by using only a pointing gesture and a reference term. Figure 5 provides a graphic definition of Limit Distance. It shows that the listener cannot distinguish the indicated object if the edge of another object intrudes into the scope of θ_P from the indicated direction. In other words, Limit Distance d_{LIM} is the distance that includes scope θ_P and distance *S* from the object's center to its edge. The difficulty of distinguishing the indicated object from other objects is dependent on the distance between one object and another object, the distance from the speaker to the objects, and the size of the object.

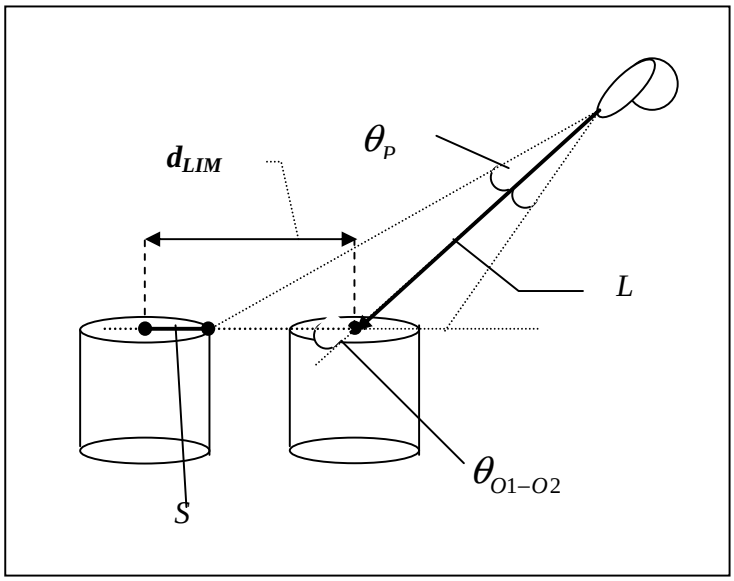


Fig.5: Definition of the Limit Distance

3.3 Additional adjectives to identify the indicated object:

When objects are close together, we identify the indicated object by using one of its properties as an adjective plus a reference term. In this research, we adopted this method to identify indicated objects when these objects fall within the Limit Distance. To identify the indicated object, one of its properties should be different from the others. The speaker may find several properties in one object, such as shape, color, and size. Using a property or set of properties different from the others, he can indicate to the listener which object is under consideration. An example of object property selection is shown in Fig. 6, where there are two pens in the environment. Because they are close together, the listener cannot identify one from the other simply by a pointing gesture and a reference term. Pen A has the following properties: Shape: STICK; Color: WHITE, and Size: SMALL. Pen B has the following properties: Shape: STICK, Color: BLACK; and Size: SMALL. In this case, the property divides the pens is Color. If the speaker wants to indicate Pen B to the listener, he identifies it using 'That BLACK one'.

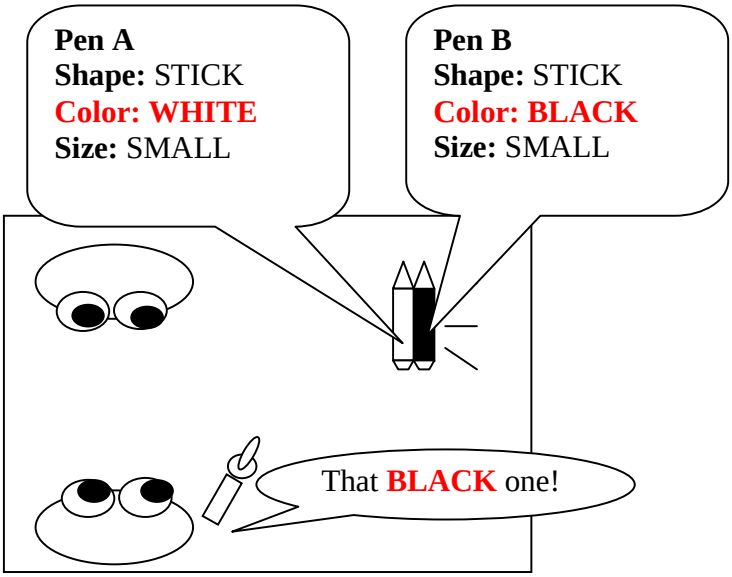


Fig.6: An example of the object property selection

4 THREE-LAYER ATTENTION-DRAWING MODEL

4.1 An overview of the proposed model

Based on the analysis of human behavior, we propose a three-layer attention-drawing model for humanoid robots using pointing gestures and verbal cues as shown in Fig. 7. The model consists of three sub-models: a Reference Term Model, a Limit Distance Model, and an Object Property Model. By selecting appropriate verbal cues based on these sub-models, the model selects the most appropriate verbal cue to indicate an object. The sub-model roles are as follows:

Reference Term Model
: To select an appropriate reference term that corresponds to the target object.

Limit Distance Model
: To identify whether using a pointing gesture and a reference term will be sufficient to indicate the target object.

Object Property Model
: To select an appropriate object's property to be used in addition to the reference term
It will be used when the robot cannot indicate the target object only by a pointing gesture and a reference term.

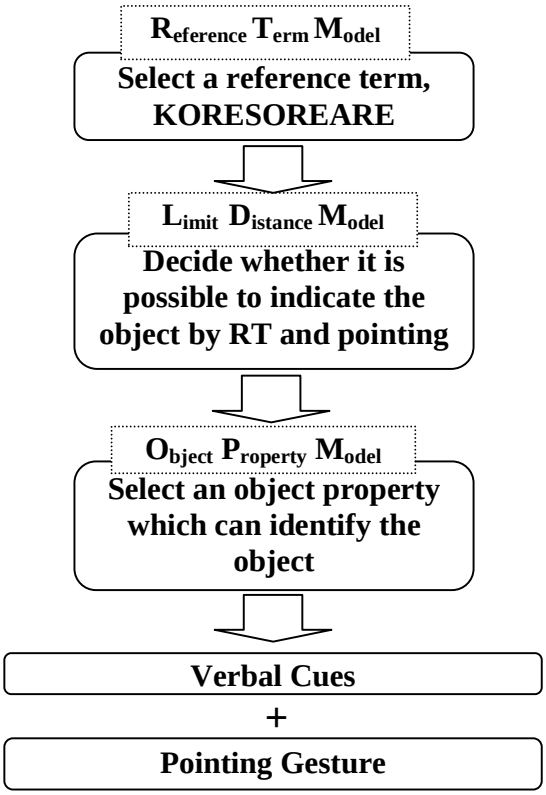


Fig.7: the Model Configuration

4.2 RTM: Reference Term Model

The RTM determines an appropriate reference term to identify the indicated object based on the approximate curve functions of reference term borders in Fig. 4. The functions are based on ellipses. Using least-squares approximations, we determined appropriate curve functions for each reference term border.
The equation of an ellipse in polar coordinates is given by:

$$r = \sqrt{\frac{a^2 b^2}{a^2 \sin^2\theta + b^2 \cos^2\theta}} \qquad \textbf{(3-1)}$$

Based on that formula, we made an approximate curve function for the *kore-sore* border as follows:

$$f_{KS}(d_{SL}, \theta_{SO}) = \begin{cases} \sqrt{\dfrac{{a_{SF}}^2 {b_S}^2}{{a_{SF}}^2 \sin^2\theta_{SO} + {b_S}^2 \cos^2\theta_{SO}}} (\theta_{SO} \le 90) & \textbf{(3-2-1)} \\ \sqrt{\dfrac{{a_{SB}}^2 {b_S}^2}{{a_{SB}}^2 \sin^2\theta_{SO} + {b_S}^2 \cos^2\theta_{SO}}} (\theta_{SO} \ge 90) & \textbf{(3-2-2)} \end{cases}$$

$$\begin{cases} a_{SF} = 0.50 \times d_{SL} + 0.13 \\ a_{SB} = 1.00 \\ b_S = 1.30 \end{cases} \qquad \textbf{(3-3)}$$

The parameters of this function are illustrated in Fig.8.

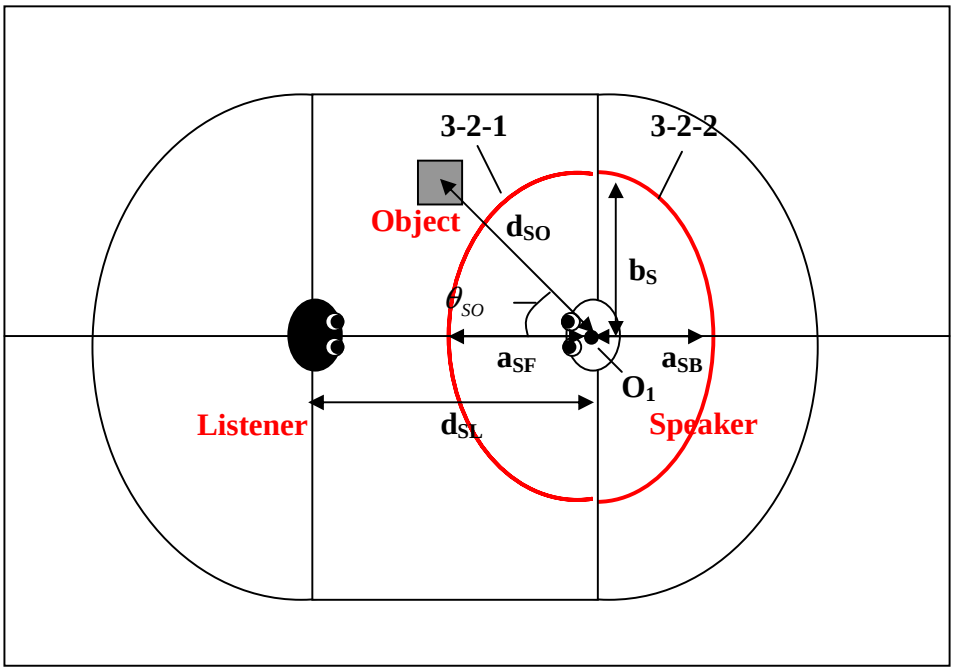


Fig.8: Parameters of an approximate curve function of the *kore-sore* border

The approximate curve function for the *kore-sore* border is the equation in the speaker's coordinate system with origin O_1 (Fig. 8). Equation (3-2) is divided into two parts. The first one is the half-ellipse in front of the speaker, and the other one is the half-ellipse behind speaker. Each ellipse has a different short axis, a_{SF}, a_{SB} (SF: Speaker Front, SB: Speaker Back), but the same long axis, b_S.

On the other hand, the approximate curve function for the *are-sore* border is given by,

$$f_{SA}(d_{SL},\theta_{SO},\theta_{LO},\theta_L)=\begin{cases}\sqrt{\dfrac{a_{SB}^{\,2}b_{SB}^{\,2}}{a_{SB}^{\,2}\sin^2\theta_{SO}+b_{SB}^{\,2}\cos^2\theta_{SO}}}\ (\theta_{SO}\ge 90, d_{SO}<d_{LO}) & \textbf{(3-4-1)}\\ \dfrac{p}{\cos(\theta_{SO,LO}-\alpha)}\ (\theta_{SO,LO}\le 90) & \textbf{(3-4-2)}\\ \sqrt{\dfrac{a_{LB}^{\,2}b_{LB}^{\,2}}{a_{LB}^{\,2}\sin^2\theta_{LO}+b_{LB}^{\,2}\cos^2\theta_{LO}}}+c_L\times\sin\theta_L\ (\theta_{LO}\ge 90, d_{LO}<d_{SO}) & \textbf{(3-4-3)}\end{cases}$$

$$\begin{cases}a_{SB}=1.25\\ b_{SB}=2.00\\ a_{LB}=0.13\times d_{SL}+1.00\\ b_{LB}=0.13\times d_{SL}+1.63\\ c_L=0.38\end{cases}\qquad \textbf{(3-5)}$$

$$\begin{cases}\alpha=\arctan\left(\dfrac{d_{SL}}{b_{LB}-b_{SB}}\right)\\ p=\dfrac{b_X\times d_{SL}}{\sqrt{d_{SL}^{\,2}+(b_{SB}-b_{LB})^2}}\end{cases},\ b_X=\begin{cases}b_{SB}(d_{SO}<d_{LO})\\ b_{LB}(d_{LO}<d_{SO})\end{cases}\qquad \textbf{(3-6)}$$

The parameters of an approximate curve function of ARE and SORE border is shown in Fig.8.

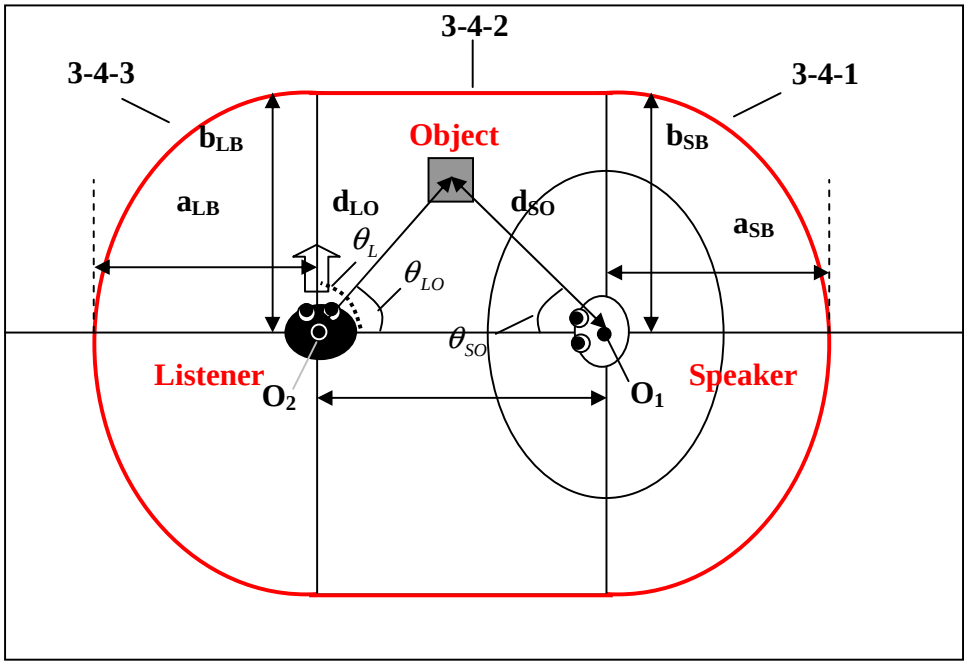


Fig.9: Parameters of an approximate curve function of ARE and SORE border

The approximate curve function for the *are-sore* border has two different coordinate systems and is divided into three parts in order to make its expression simple. The first coordinate system is that of the speaker, with origin O_1. The second is the listener coordinate system, with origin O_2. If the distance between the speaker and the object, d_{SO}, is shorter than the distance between the listener and the object d_{LO} (the object is closer to the speaker), we use the equation in the speaker coordinates. On the other hand, if d_{LO} is shorter than d_{SO} (the object is closer to the listener), we use the equation of listener coordinates. Equation (3-3-1) is the equation representing the half-ellipse behind the speaker in speaker coordinates, and has short axis a_{SB} and long axis b_{SB} as its parameters. Equation (3-3-3) is that representing the ellipse behind the listener in listener coordinates, and has short axis a_{LB} and long axis b_{LB} as its parameters. It includes an additional term $c_L \times \sin\theta_L$, which is dependent on the listener's direction θ_L. Equation (3-3-2) is the one representing the line between the half-ellipse behind the speaker and the half-ellipse behind the listener. Parameters α, p are the parameters of this line, and have different expressions corresponding to the two coordinate systems. The parameters of an approximate curve function of the *are-sore* are shown in Fig. 9.

The RTM uses the following rules to decide the most appropriate reference term to use.

1. If the object is closer to the speaker:
   A) If $d_{SO} \le f_{KS}(d_{SL}, \theta_{SO})$, use the reference term *kore*.
   B) If $d_{SO} > f_{KS}(d_{SL}, \theta_{SO})$ and $d_{SO} \le f_{SA}(d_{SL}, \theta_{SO}, \theta_{LO}, \theta_L)$, use the reference term *sore*.
   C) If $d_{SO} > f_{SA}(d_{SL}, \theta_{SO}, \theta_{LO}, \theta_L)$, use the reference term *are*.

2. If the object is closer to the listener:
   A) If $d_{LO} \le f_{SA}(d_{SL}, \theta_{SO}, \theta_{LO}, \theta_L)$, use the reference term *sore*.
   B) If $d_{LO} > f_{SA}(d_{SL}, \theta_{SO}, \theta_{LO}, \theta_L)$, use the reference term *are*.

4.3 LDM: Limit Distance Model

The LDM estimates whether the listener can identify the indicated object by using only a pointing gesture and a reference term, based on Limit Distance. Limit Distance d_{LIM} is given by

$$d_{LIM} = f(S, L, \theta_P) = \frac{\tan\theta_P \times L}{\sin\theta_{O1-O2} - \tan\theta_P \cos\theta_{O1-O2}} + S ,$$

The parameters for which are given in Fig. 5. The LDM uses the following rules to determine whether the model needs to use the properties of an object:

1. Use the object's properties when $d_{LIM} \leq f(S, L, \theta_P)$.

2. Do not use them when $d_{LIM} > f(S, L, \theta_P)$.

4.4 OPM: Object Property Model

The OPM chooses a property of the indicated object that is different from the other objects when these objects are all within the Limit Distance. The OPM contains a list of the properties of each object, and it finds an appropriate property to use by comparing each property among the objects. In this research, however, the OPM has only color as an object property. A system for determining object properties and an algorithm to compare them are under consideration.

5 DEVELOPMENT OF THE THREE LAYER ATTENTION-DRAWING MODEL

5.1 Hardware Configuration

We implemented the three-layer attention-drawing model to a communication robot, Robovie (Kanda et al. 2004). Robovie (Fig. 10) is a humanoid robot that has a head, two arms, a body, and a wheeled-type mobile base, to communicate with humans. On the head it has two CCD cameras as eyes and a speaker as a mouth. The speaker can output recorded sound files installed on the internal control PC in the body. Its height and weight are 120 cm and 40 kg, respectively. Degrees of freedom (DOFs) of the robot are as follows: two DOFs for the wheels, three DOFs for its neck, and four DOFs for each arm.

In this research, we used a motion-capturing system* to bypass the technical difficulty of implementing visual recognition functions for humans' posture and the positions of objects. The motion-capturing system (Fig. 11) allows us to obtain accurate 3D positions from markers attached to Robovie, a subject, and the objects. The robot and the motion-capturing system are connected via a 100-Mbps Ethernet.

* http://www.vicon.com/

Fig. 10: Robovie

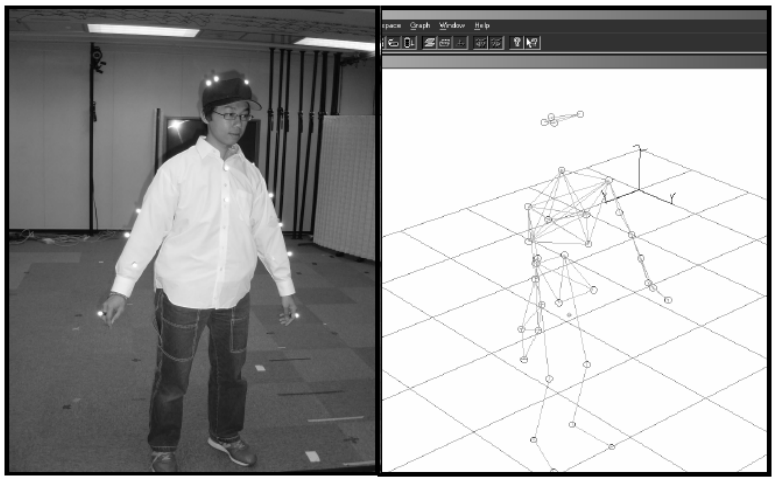


Fig.11: Motion capturing system (Vicon)

5.2 System Configuration

Figure 12 shows the configuration of the developed system. The roles of each module in Fig. 10 are as follows:

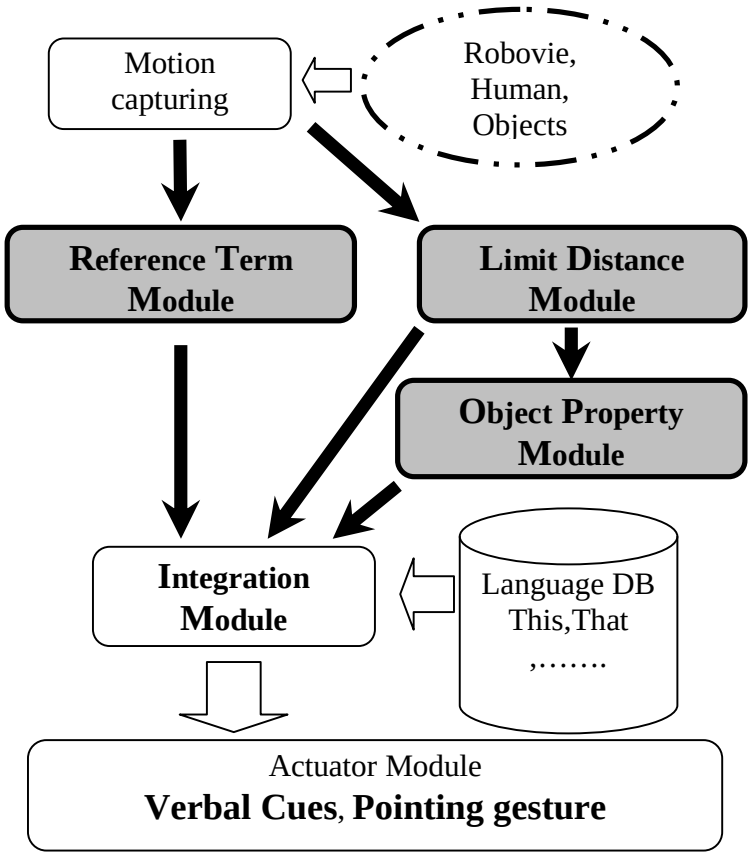


Fig.12: System Configuration

5.2.1 Reference Term Module

The Reference Term Module is the implementation of the Reference Term Model. This module receives the 3D positions of markers attached to Robovie, a subject, and objects from a motion capturing system. Based on the relationships among the marker positions, it determines which reference term should be used. The module outputs its decision to the Integration module.

5.2.2 Limit Distance Module

The Limit Distance Module is the implementation of the Limit Distance Model. Figure 13 illustrates how the module operates. The module first receives the 3D positions of markers attached to Robovie and the objects from the motion-capturing system as input, and decides whether to use the properties of the objects to identify the indicated object, based on Limit Distance Model. This module outputs the decision to both the Object Property Module and the Integration Module.

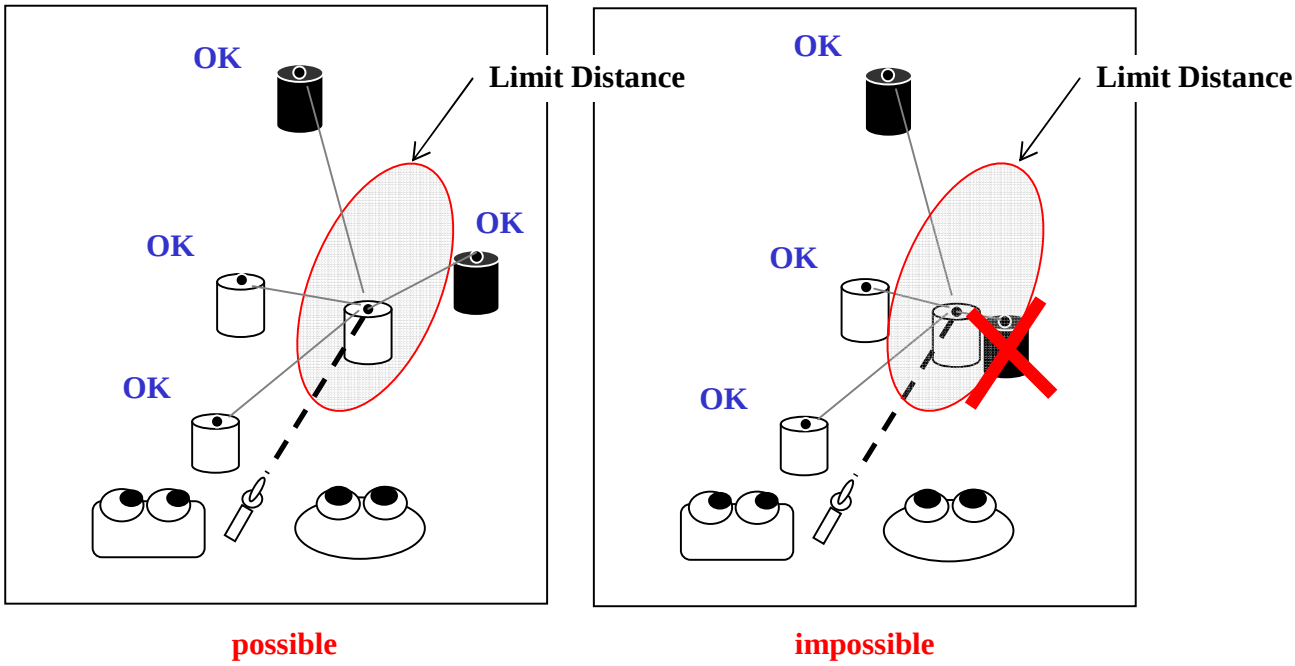


Fig. 13: The decision in the Limit Distance Module

5.2.3 Object Property Module

The Object Property Module is the implementation of the Object Property Model. If the Limit Distance Module determines that a certain property of an object be used to identify the object, it selects the most appropriate property of that object and identifies the object according to the Object Property Model. It then outputs its decision to the Integration Module. In this research, we only use one property, color, as an object property to identify the indicated object.

5.2.4 Integration Module:

The Integration Module integrates the decision of each sub-model module and chooses appropriate verbal cues from the language database. It outputs the result in order to make the robot speak the verbal cue.

5.2.5 Actuator Module

The Actuator Module controls the robot's hardware to make pointing gestures and verbal cues based on the command from the Integration Module. It moves the robot's arm in the direction of the target object to produce the pointing gesture. If the target object is on the right side of the robot, the right hand is used; otherwise it uses the left hand.

6. EXPERIMENTS

To confirm the effectiveness of the developed system, we conducted the following three experiments to:
1. Verify of the effectiveness of the RTM.
2. Verify of the effectiveness of the LDM.
3. Verify of the effectiveness of the system.

6.1 Experiment to verify the effectiveness of the RTM

6.1.1 Hypothesis

We tested the hypothesis that with RTM the system can choose an appropriate reference term for subjects to understand the object's location.

6.1.2 Experimental Procedure

[Brief Overview]

We verified the RTM's effectiveness in this experiment. Every three seconds Robovie repeated a reference term selected by the RTM that indicated a round object (11 cm in radius and 30 cm tall) moved at a constant speed by an experimenter. Subjects evaluated each spoken reference term. This experiment consisted of the following three sessions to verify the RTM effectiveness:

- A session in which the object is moved in direction (A) a distance of around 4.0 m:
  Investigate the usage of reference terms when the object position is moved from a place close to the speaker to one close to the listener
- A session in which the object is moved in direction (B) a distance of around 3.0 m:
  Investigate the usage of reference terms when the object position is moved from a place close to the speaker to one far from the speaker
- A session in which the object is moved in direction (C) a distance of around 3.0 m:
  Investigate the usage of reference terms when the object position is moved from a place close to the listener to one far from the listener

To minimize the influence of the object's movement, we slowed down the object speed to around 10 cm/s.

[Subjects]

The subjects were thirteen male & female university students, ten of whom had participated in the experiment in Section 6.2.

[Experiment Procedure]

The experiments included Robovie, which repeated a reference term every three seconds, a subject, and an object moved in one direction. The experiment consisted of three sessions, and in every session we changed the direction from (A) to (C) (Fig. 14) in which we physically moved the object.

[Verification Method]

Subjects evaluated the acceptability of each reference term by tri-level evaluation: 'Very Acceptable', 'Acceptable', or 'Unacceptable'.

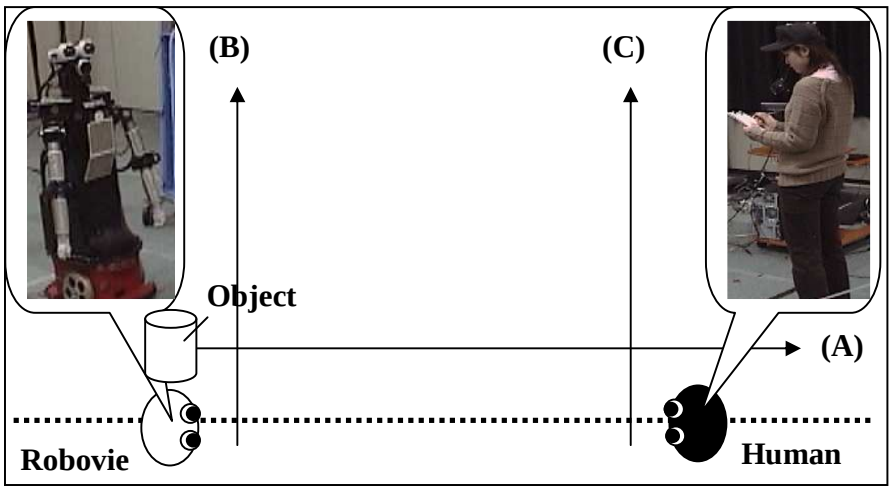


Fig.14: Object moving directions in the verification of the RTM

6.1.3 Experimental Results

[Prediction]

If the hypothesis is correct, subjects will not feel incongruity to the reference terms used by the robot, which is determined with the RTM. Thus, the subject will evaluate the reference terms as 'Very Acceptable' or 'Acceptable'.

[Verification]

Table I shows the results from the questionnaire. The average rate of choice for 'Very Acceptable' and 'Acceptable' was 96.02%. Thus, based on the RTM, the robot used an appropriate reference term so that subjects were able to identify which object was pointed out by the robot. Although there was a 3.97% response for 'Unacceptable', we believe that this result is within an acceptable range because individual differences in the use of reference terms are diverse to a certain extent.

Table I: Average rate of subjects' answer to each category in the questionnaire

Evaluation	Average rate
Very acceptable	81.95%
Acceptable	14.07%
Unacceptable	3.97%

6.2 Experiment to verify the effectiveness of the LDM

6.2.1 Hypothesis

We tested the hypothesis that with the LDM the system can identify the limits of the area where people can identify an object using only a pointing gesture and verbal cues.

6.2.2 Experimental Procedure

[Brief Overview]

We verified effectiveness of the LDM in this experiment, in which Robovie indicated to a subject one of two round objects (both are 11cm radius and 30cm tall) in the environment. The distance between the two objects

was changed as shown in Fig. 15. Subjects responded which object they thought the robot was indicating. Through this evaluation, we would like to investigate whether the subject can correctly identify the object within the areas certified by the LDM.

[Subjects]

The subjects were ten male & female university students, all of who had taken part in the experiment in Section 6.1. There should not be any influence on the experiment in Section 6-1 and this experiment. We conducted these experiments at the same time.

[Experiment Procedure]

The experiment comprised two sessions. In Session 1, Object 1 was placed on the black circle in the positions marked for Session 1 in Fig. 15. Object 2 was placed alongside Object 1. In each trial, Object 2 was placed 30 cm from its previous position in direction (a). When the three trials were finished, Object 2 was next placed 30 cm from Object 1 in direction (b), and the same steps were repeated until the three trials were finished. When all the trials were finished in Session 1, Object 1 was placed in the black circle in the positions for Session 2 in Fig. 15, and the same process was applied to the trials in Session 2. In each trial, Robovie indicated to a subject one of the two objects. The indicated object was decided randomly and subjects decided which object was the indicated one.

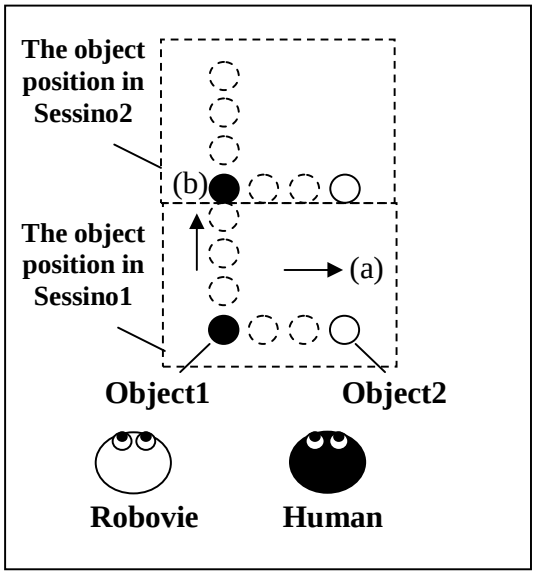


Fig.15: Experimental environment in the verification of the LDM

[Verifying Method]

Subjects responded which object they thought the robot was indicating. We measured the recognition rate of the indicated object, which is calculated by 'the number of correct answer of subjects / the number of indication by the robot.'

6.2.3 Experimental Results

[Prediction]

If the hypothesis is correct, when the distance between the two objects falls within the Limit Distance, subjects will find difficulty in identifying the object indicated by the robot. Thus, the recognition rate will be lower in the 'Area Not certified by LDM' than the one in the 'Area Certified by LDM'.

[Verification]
The recognition rate in the area certified by the LDM was 83.3%, while the rate outside the certified area was 56.6% (Table II). Consequently, with the LDM the system is able to identify the area in which people can identify the indicated object with a pointing gesture and verbal cues.

Table II: Recognition rate between an area certified and not certified by LDM

Area	Recognition Rate
Area Certified by LDM	83.3%
Area Not certified by LDM	56.6%

6.3 Experiment to verify the system's effectiveness

6.3.1 Hypothesis
In this experiment (Fig.16), we tested the following two hypotheses.
Hypothesis 1: The system can correctly draw the listener's attention to the one specific object among several objects in an environment
Hypothesis 2: Using reference terms, the system can quickly draw the listener's attention to the specific object among several objects in an environment.

6.3.2 Experimental Procedure
[Brief Overview]
In this experiment, illustrated in Fig. 16, we verified the effectiveness of the attention-drawing system. Robovie indicated to subjects one of five round objects (all objects were 11 cm in radius and 30 cm tall) in the environment. The verbal cues that the system used had two variations:

1. Verbal cues with a reference term, which is decided by our proposed three-layer model, utilizing a pointing gesture.
2. Verbal cues with three kinds of character written on the object (numeric, alphabetical, and Japanese) that can identify each object, utilizing a pointing gesture.

The second cues were set up for a comparison of the effectiveness of reference terms and other verbal cues. Subjects responded with the object they thought the system indicated.
[Subjects]
The subjects were 21 male and female university students, all of whom were different from those who participated in the experiments in Sections 6.1 and 6.2.
[Verifying method]
We verified Hypothesis 1 using the recognition rate of the indicated object and Hypothesis 2 by using questionnaire data compiled after the experiment. The questionnaire entries were as follows:

1. With which can you more **quickly** recognize the indicated object, reference terms or characters?
2. With which can you more **accurately** recognize the indicated object, reference terms or characters?
3. Which seem to be more **natural,** reference terms or characters?

Subjects chose from the following three answers:

1. Reference terms;
2. Characters;
3. No preference.

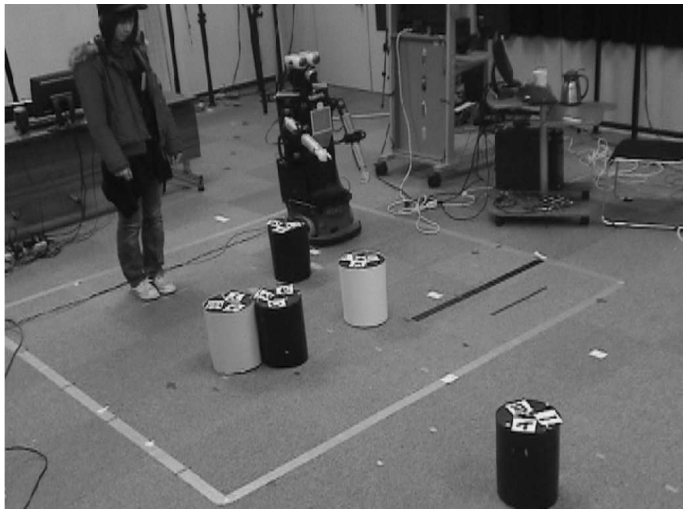


Fig. 16: Experiment to verify the system's effectiveness

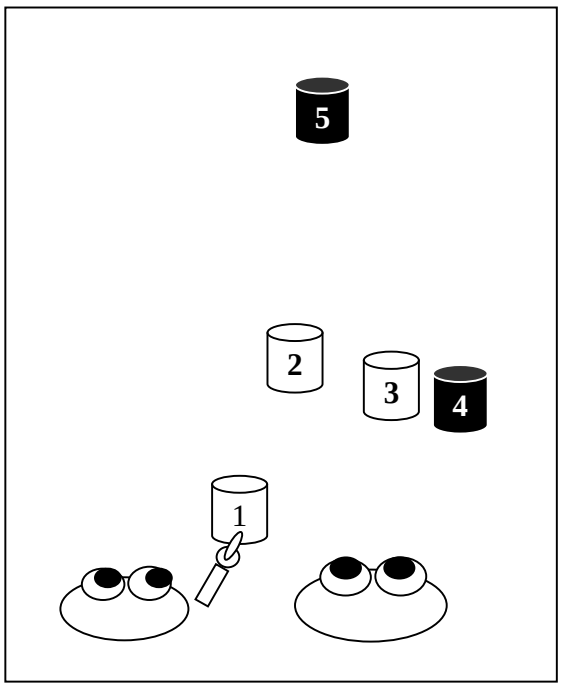


Fig. 17: Experimental Environment to verify the system's effectiveness

6.3.3 Experimental Results

[Prediction]

If the hypothesis 1 is correct, subjects will be able to identify the object indicated with verbal cues determined by the system. Thus, the recognition rate of the object indicated with the verbal cues will be as high as the one with characters; we assume that subject can accurately identify the object when the system uses the characters to indicate the object.

If the hypothesis 2 is correct, subjects will feel more quick to find the object indicated with the reference terms than the object indicated with the characters. Thus, the subject will tend to choose the reference term for the question of 'quickness'.

[Verification of hypothesis 1]
Table III shows the recognition rate of the indicated objects. Both the recognition rate using reference terms and that using characters are almost the same. The recognition rate results show that both a character with pointing and a reference term with pointing could indicate the object in the environment. Consequently, the system can correctly draw listener's attention to the indicated object among multiple objects in the environment.

[Verification of hypothesis 2]
Results for questionnaire are shown in Fig. 18. A chi-square test indicates that the number of subjects who answered 'reference term' was significantly high ($\chi^2_{(2)} = 14.000, p < .01$; 'reference term' > 'character' : $p < .05$). Thus, it is proved that the system can quickly draw listener's attention to the indicated object by using reference terms.

Table III: Recognition Rate of indicated object in each condition

Utterance	Recognition Rate
Verbal cues with reference term utilizing pointing gesture	93.33%
Verbal cues with characters utilizing pointing gesture	92.38%

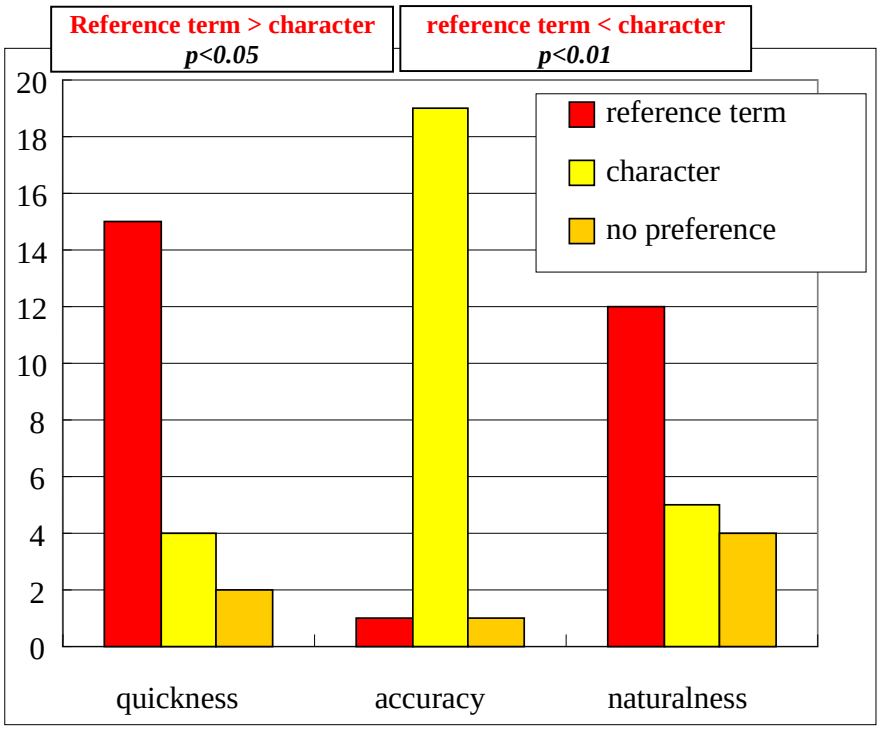


Fig. 18: Questionnaire data in verification of the system's effectiveness

Meanwhile, for accuracy, the number of subjects who answered 'characters' was significantly high ($\chi^2_{(2)} = 38.853, p < 0.01$; 'reference term' < 'character' : $p < .01$). Thus, there seems to be some inconsistency in the result for the recognition rate and the questionnaire answer relating to accuracy. Subjects felt that the robot's indication with the character is rather more accurate than that with the reference term. However, they were able to identify the indicated object with enough accuracy (93.33% with the reference term; 92.38% with the

character). There was no significant difference in the naturalness, but at least the robot's indication with the reference term is not more unnatural than indication with the character.

7. DISCUSSION

7.1 Summary of the results

The developed three-layer attention-drawing model consists of these sub-models: the Reference Term Model (RTM), the Limit Distance Model (LDM), and the Object Property Model (OPM). The robot selected an appropriate verbal cue to use based on this three-layer model. Experiments were conducted to verify the effectiveness of each of the three sub-models as well as that of the whole three-layer attention-drawing model. The results revealed the following points:

1. Effectiveness of the RTM (Experiment 6.1)
   In the experiment, most of the reference terms the robot used were evaluated as being within an understandable range ('Quite Acceptable' or 'Acceptable'). Thus, the robot was able to choose an appropriate reference term based on the RTM.
2. Effectiveness of the LDM (Experiment 6.2)
   The experimental result revealed that the LDM was able to identify the area where people can understand the robot's pointing gesture. This means the robot was able to identify a situation in which it should not use a simple expression.
3. Effectiveness of the three-layer attention-drawing model (Experiment 6.3)
   In the experiment, the robot used a pointing gesture and verbal cues, which were generated based on the model. As a result, subjects responded that they were able to quickly identify the object indicated by the robot with the three-layer attention-drawing model. Moreover, they accurately identified the objects indicated by the robot both with and without the three-layer attention-drawing model, although they answered in the questionnaires that indication by the robot without the model was more accurate. One reason why the subjects answered the characters are more accurate is considered to be the issue of the subjects' own interpretations. That is, since subjects consider that the character is more specific than a gesture and a reference term, that they may feel the need to call it more accurate. (These subjective feelings did not correspond with the facts: the subjects' recognition rate was slightly higher for indication by the robot with the three-layer attention-drawing model). We believe that this is similar to what happens in humans' casual conversations: reference terms used in conversation among humans, such as 'please look at this', are subjectively quick but inaccurate.

7.2 Perspectives for android science

In "android science," many research works focus on the human-likeness of robots. Particularly, appearance and simple motions, such as eyeball motion, blinking, etc., should be among the main focuses of android research. We believe that the combination of these fundamental techniques will soon result in human-like interaction between people and an android robot. For example, we are looking forward to the development, possibly in the near future, of an android robot that can sit down beside a person as if it were a real human. The next step will be to acquire human-like daily conversation.

The research presented in this paper revealed that the three-layer attention-drawing model has the potential to make a humanoid robot capable of such human-like conversation about objects with reference terms and gestures, such as 'look at this', and 'please pick up this box'.

One of the most important future works in android science will be to identify the effect of combining a human-like appearance (such as the one realized in a very human-like android (Ishiguro 2005)), and that of the conversation-level achieved in this research. We can expect that an android will be able to perform more human-like interaction with people based on such integration. For example, people may feel communication with an android robot to be more believable and natural. On the other hand, if an android robot does not use such reference terms in conversation, people may perceive its way of speaking as too robotic. This may result in a more unnatural impression of it.

7.3 Generality and Limitation

Since the experiments were only with one particular robot, Robovie, the generality of these results with respect to other robots is limited. In other words, we cannot be sure whether the findings from these experiments can be applied to all other humanoid robots and android robots. We believe, however, that it is a realistic enough setting and a good start for research on the gestures and verbal cues given by humanoid robots. Moreover, although Robovie has a less sophisticated design than other humanoid robots such as Honda Corp.'s Asimo, the experiments revealed the effectiveness of the developed system. Thus, we believe that the developed model is probably applicable for other humanoid robots and android robots that have a similarly simple or better appearance. Another limitation derives from the simplified settings of the experiments. To use the model in our daily environment, we should also take into account people's movements when they are acting, walking, and so forth. We believe that we can address these "scaling-up" issues by applying a more developed model to more realistic settings, such as a field trial in a science museum (Shiomi, 2006) or a simulated environment in a laboratory, which should be one of our future works.

7.4 How to take into account of appearance of android robots

To apply the model to an android robot, we should take into account how human-like in appearance it is. The developed model is mainly related to the processing of information based on the positional relationships among people, a robot, and objects. This work did not focus on the appearance problem. One important contribution that we can expect from human-like appearance is toward natural and believable feelings, as we discussed in the previous section.
Moreover, there will be a greater expectation toward robots having a more human-like appearance. This expectation is considered to result from the higher standards of robot performance that people are continually demanding. Because this expectation for more human-like appearance probably affects the robot's pointing gesture, we may need to adjust the Limit Distance Model (LDM). The effect of the Robovie-type robot's pointing gesture was relatively weak, leading us to adjust the limit distance in the LDM to 10 degrees. Of course, this parameter is hardware-dependent; Robovie only had one pointing finger on a spherical hand. Several current humanoid robots feature five-fingered hands, however, and there are also android robots equipped with human-like hands with five fingers. We believe that such a sophisticated hand will have a more powerful effect in pointing gestures. Consequently, we will be able to use a smaller limit distance in the LDM.

7.5 How to take into account of a different language

The Reference Term Model (RTM) is language-dependent. In Japanese, there are three reference terms: *kore*, *sore*, and *are*. The usage borders among these reference terms are associated with the positions of speaker, listener, and the object. In English, for example, there are two main reference terms, 'this' and 'that' (as well as 'over there'), whose usage border is only associated with the position of the speaker and the object. We believe that such a dependency can be implemented in the RTM. In other words, the developed attention-

drawing model is probably capable of other operating in other languages by switching the current RTM to one for a different language. It should be one of the interesting future works to validate its effectiveness in cross-cultural or bilingual environment.

8. CONCLUSION

We have developed a three-layer attention-drawing model based on observations of human conversations that refer to objects in the environment. The first layer, the RTM, associates positional relationships with reference terms. The second layer, the LDM, associates the pointing gesture with the reference term selected by the RTM. The third layer, the OPM, associates other supplemental verbal cues with the reference term. The model was implemented to a humanoid robot, Robovie. The experimental results revealed that the robot could quickly and correctly indicate the target object with a pointing gesture and reference terms. Thus, the three-layer attention-drawing model enables a robot to exploit its human-like bodily expression and verbal cues for human-like casual conversation. We believe that this capability is one of the essential components for making humanoid and android robots fully human-like in communication with humans.

9. Acknowledgement

This research was supported by the Ministry of Internal Affairs and Communications of Japan.

References

C. Breazeal and B. Scassellati: 'Infant-like social interactions between a robot and a human caretaker.' *Adaptive Behavior, 8(1)*, 2000.

J. Goetz, S. Kiesler, and A. Powers.: 'Matching robot appearance and behavior to tasks to improve human-robot cooperation.' *Proceedings of the 12th IEEE Workshop on Robot and Human Interactive Communication. RO-MAN 2003*, Oct. 31 - Nov. 2, 2003

M. Imai, K. Hiraki and T. Miyasato: 'Physical Constraints on Human Robot Interaction', *Proceedings of 16th International Joint Conference on Artificial Intelligence (IJCAI99), Vol.2*, PP.1124--1130, 1999

M. Imai, T. Ono and H. Ishiguro: "Physical Relation and Expression: Joint Attention for Human-Robot Interaction", *IEEE Transactions on Industrial Electronics, Vol. 50, No. 4, ITIED 6*, pp.636-643, 2003

H. Ishiguro: 'Android Science – Toward a new cross-interdisciplinary frame works –', *CogSci-2005 Workshop*, pp 1-6

M. Kamasima, T. Kanda, M. Imai, Tetuo Ono, Daisuke Sakamoto, Hiroshi Ishiguro, and Yuichiro Anzai: 'Embodied Cooperative Behaviors by an Autonomous Humanoid Robot', *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2004)*, pp.2506-2513, 2004.

T. Kanda, H. Ishiguro, M. Imai and T. Ono: "Development and Evaluation of Interactive Humanoid Robots", *Proceedings of the IEEE* Vol.92, No.11, pp. 1839-1850, 2004.

T. Kanda, T. Miyashita, T. Osada, Y. Haikawa and H. Ishiguro: 'Analysis of Humanoid Appearances in Human-Robot Interaction', *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2005)*, pp. 62-69, 2005.

C. Kidd and C. Breazeal: ‘Effect of a Robot on User Perceptions’, *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2004)*, 2004.

C. Moore and Philip J. Dunham eds: ‘Joint Attention: Its Origins and Role in Development’, *Lawrence Erlbaum Associates*, 1995.

K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano: ‘Real-Time Auditory and Visual Multiple-Object Tracking for Robots’, *Proc. Int. Joint Conf. on Artificial Intelligence*, pp.1425-1432, 2001.

B. Scassellati: ‘Investigating Models of Social Development Using a Humanoid Robot’, *Biorobotics, MIT Press*, 2000.

D. Sakamoto, T. Kanda, T. Ono, M. Kamashima, M. Imai, and H. Ishiguro: ‘Cooperative embodied communication emerged by interactive humanoid robots’, *International Journal of Human-Computer Studies*, Vol. 62, pp. 247-265, 2005.

J. G. Trafton, N.L. Cassimatis, M. Bugajska, D. Brock, F. Mintz and A. Schultz: ‘Enabling effective human-robot interaction using perspective-taking in robots.’ *IEEE Transactions on Systems, Man and Cybernetics*, pp. 460-470. Volume 25. Issue 4, 2005.

D. McNeill, ‘PSYCHOLINGUISTICS : A NEW APPROACH’, ISBN 4-7819-0593-5, PRINTED IN JAPAN, 1990

M. L. Walters, K. Dautenhahn, K. L. Koay, C. Kaouri, R. te Boekhorst, C. L. Nehaniv, I. Werry, and D. Lee: ‘Close encounters: Spatial distances between people and a robot of mechanistic appearance.’, *Proc. IEEE-RAS International Conference on Humanoid Robots (Humanoids2005)*, pp. 450-455, 2005

H. Kozima and E. Vatikiotis-Bateso: “Communicative criteria for processing time/space-varying information,” Proc. 10th IEEE International Workshop on Robot and Human Communication, IEEE, pp. 377-382.

Y. Nagai: “Learning to Comprehend Deictic Gestures in Robots and Human Infants,” In Proceedings of the 14th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN'05), pp. 217-222, August 2005.

T. Mizuno, Y. Takeuchi, H. Kudo, T. Matsumoto, N. Onishi, and T. Yamamura: “Informing a Robot of Object Location with Both Hand-Gesture and Verbal Cues,” IEEJ Trans. EIS, Vol. 123, No. 12, 2003 (Japanese)

A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer: “A multi-modal object attention system for a mobile robot,” IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2005), pp. 1499-1504, 2005.

Z. M. Hanafiah, C. Yamazaki, A. Nakamura and Y. Kuno: “Understanding Inexplicit Utterances Using Vision for Helper Robots,” Proceedings of the 17th International Conference on Pattern Recognition, /CD-ROM V44_2_04.pdf, Cambridge, UK, August 23-26, 2004.

T. Inamura, M. Inaba, and H. Inoue: “PEXIS: Probabilistic Experience Representation Based Adaptive Interaction System for Personal Robots,” Systems and Computers in Japan, Vol. 35, No. 6, pp. 98--109, 2004.

B. Scassellati: “Investigating Models of Social Development Using a Humanoid Robot,” Biorobotics, MIT Press, 2000.

M. Shiomi, T. Kanda, H.Ishiguro, N. Hagita, ‘Interactive Humanoid Robots for a Science Museum’, ACM 1st Annual Conference on Human-Robot Interaction (HRI2006), pp. 305-312, 2006.