It's Not Polite to Point

Generating Socially-Appropriate Deictic Behaviors Towards People

Phoebe Liu[1,2], Dylan F. Glas[1,2], Takayuki Kanda[1], Hiroshi Ishiguro[1,2], Norihiro Hagita[1]

(1)Intelligent Robotics and Communication Laboratories
Advanced Telecommunications Research Institute International
2-2-2 Hikaridai, Keihanna, Kyoto, Japan
{phoebe, dylan, kanda, hagita}@atr.jp

(2)Faculty of Engineering Science
Osaka University
1-3 Machikaneyama, Toyonaka, Osaka, Japan
ishiguro@sys.es.osaka-u.ac.jp

***Abstract*— Pointing behaviors are used for referring to objects and people in everyday interactions, but the behaviors used for referring to objects are not necessarily polite or socially appropriate for referring to humans. In this study, we confirm that although people would point precisely to an object to indicate where it is, they were hesitant to do so when pointing to another person. We propose a model for generating socially-appropriate deictic behaviors in a robot. The model is based on balancing two factors: understandability and social appropriateness. In an experiment with a robot in a shopping mall, we found that the robot's deictic behavior was perceived as more polite, more natural, and better overall when using our model, compared with a model considering understandability alone.**

***Index Terms*—human-robot interaction; social robots; pointing gesture**

I. Introduction

People often use pointing gestures to objects and people in everyday interactions. However, there are important differences in the way we gesture towards objects and the way we gesture towards people. For example, if a person saw a certain badly designed cellphone model, he would generally tell his friend how ugly or bulky this cellphone was while pointing directly at it. On the other hand, if he saw someone with a bad fashion sense and chose to make snide comments about that person's choice of clothing, he would probably discreetly point out that person to his friend, using a subtle pointing gesture in order to avoid drawing attention to himself and to spare hurting the ill-dressed individual's feelings. These two situations illustrate a fundamental difference between pointing to people and pointing to objects (Figure 1).

When pointing to people, it is important for the speaker to consider the possibility that the referent may become aware of the conversation. If an obvious deictic behavior is directed toward the referent, it may cause the referent to feel singled out, self-conscious, and uncomfortable. If this happens, the speaker has created a socially-awkward situation. Hence, when pointing to a person, it is important to consider the social appropriateness of the gesture, something that is not a factor when pointing to objects.

In a "closed" conversation, where the speaker does not intend the referent to hear the conversation, e.g. when saying something negative about the referent, the significance of social appropriateness is even more obvious. In order to avoid causing unnecessary pain or awkwardness to the referent, the speaker would be cautious not to make the referent aware of the conversation. However, in "open" conversation, e.g. when saying something neutral or good about the referent, the speaker might gesture towards the person in a more obvious way.

Figure 1: Pointing precisely to people makes them feel self-conscious

Existing models for generating deictic behaviors in robots are typically designed for referring to objects, and thus do not consider this element of social appropriateness. In this study, we present a model for generating socially-appropriate deictic behaviors for pointing to people. From a study of human behavior, we confirmed that people usually do not use precise pointing gestures, that is, they do not use index-finger to directly point towards another person, and that this phenomenon becomes even more pronounced in the case of private, or "closed," conversation.

We consider the choice of deictic behavior to involve a balance between understandability and social appropriateness: more precise pointing gestures can increase understandability, but the increased precision can be socially inappropriate. Based on this concept and the data from our human behavior observations, we have developed a model enabling a robot to reproduce human deictic behavior towards people.

Finally, we present results from an experiment conducted with a robot in a shopping mall, showing that people evaluated the robot's behaviors as more natural and polite when social appropriateness was considered in behavior selection.

II. Related Work

A. Studies of Human Pointing Behavior

According to Kendon, the intention of precise pointing is to single out an object which is to be attended to as a particular individual object [1]. He categorized this type of pointing as the Index Finger Extended, for which not only the index finger,

but almost any extensible body part or held object can be used. The idea that index finger pointing singles out a particular entity is a well-established idea in human science literature, and is also used as a basis for our categorization.

Some studies have included using reference terms for people. In such studies, the focus was mainly on generating a referring expression (i.e. "This is the coach") to single out someone as an individual person [2-4]. Accordingly, we also consider verbal descriptive terms as part of our model for generating deictic behavior.

B. Human Robot Interaction

Similar to Kendon's work of index finger pointing to single out an object, studies have attempted to model the idea of pointing as a way to resolve ambiguity. Bangester *et al.* focused on the use of full pointing (arm fully extended) and partial pointing (elbow bent) by varying the number of pictures in an array to manipulate the ambiguity of a reference [5]. We will combine this idea of resolving ambiguity with an additional politeness factor that applies when pointing to people.

Some studies in human-robot interaction have focus on generating human-like multimodal referring acts using both speech and gesture for objects [6-8], and space [9,10]. Brooks and Breazeal [11] describe a framework for multimodally referring to objects using a combination of deictic gesture, speech, and spatial knowledge. Schultz et al. focused on spatial reference for a robot using perspective taking [12]. In these studies, the robot points to a static object in the environment and produces an appropriate deictic behavior that indicates where the target is. We will also study multimodal behaviors in human-robot interaction, but with a focus on the social aspects of pointing to people.

III. DATA COLLECTION

A. Objective

We collected data from observations of real human deictic behavior so we could generate a model enabling a robot to point naturally to people. Since pointing to objects has been explored extensively in other research, we chose to focus on ways that pointing behaviors vary when pointing to people. In particular, we were interested in examining three factors:

Object vs. person: As we discussed in the introduction, we expected that people would point precisely to objects but less precisely to people.

Open vs. closed: We expected that people would use less obvious gestures in "closed" conversation, e.g. talking about someone in a negative way, than in "open" conversation.

Known vs. unknown: We wonder if people's behavior would be different if they already know the referent, such as in the case when saying name would be enough to identify the referent without ambiguity.

B. Procedure

We conducted the data collection in a shopping mall, with 17 participants (11 female, 6 male, average 23.7 years old), who were paid. We asked the participants to imagine a situation in which they are talking with a friend (role-played by a confederate). The confederate asked the participant's opinions about other visitors in the mall, and the participant freely answered using deictic behaviors.

(a) Gaze only (b) Casual pointing (c) Precise pointing

Fig. 2. Categorization of different pointing types

We measured the behavior of the participants under 5 scenarios, chosen to measure the factors described above. The scenarios were defined as follows:

- **Object**: Referring to a product in the shopping mall that does not belong to either the participant or the confederate (e.g. "Which of these cellphones do you think looks better?").
- **Open/Known**: Referring to a mutual friend (one of two other confederates) in an open conversation. (e.g. "With which of our friends did you take the same bus to the mall?")
- **Open/Unknown**: Referring to a random, unknown customer in an open conversation (e.g. "Which person did you see at the train station yesterday?")
- **Closed/Known**: Referring to a mutual friend (one of two other confederates) in a closed conversation, such as gossiping negatively. (e.g. "Which of our friends do you think has no fashion sense?")
- **Closed/Unknown**: Referring to a random, unknown customer in a closed conversation (e.g. "Which person do you think looks unfriendly?")

Each scenario consisted of 6 questions, which were counter-balanced. Video of each participant's behaviors was recorded, and as we expected that positions of surrounding people might affect the speaker's deictic behavior (i.e., identifying a referent among many customers is more difficult than among only a few customers), we used a human tracking system [13] to capture the positions of the people in the environment.

The degree of crowding could not be explicitly controlled since the experiment was conducted in a shopping mall. However, all trials were conducted under similar conditions during weekday mornings and afternoons, with an average of 10.46 people present in the environment across all trials.

For each question, the speaker's pointing type and use of a verbal descriptive term were coded and categorized from the recorded videos, as explained below.

C. Categorization of Pointing Types

We classified pointing gestures into three categories (see Fig. 2): "gaze only", "casual pointing", and "precise pointing". Gaze only was defined as when the speaker only gazes in the direction of the referent, without the use of any other pointing gestures. Casual pointing was coded as when the arm was only

TABLE I. RATIO OF BEHAVIORS PERFORMED FROM DATA COLLECTION

Scenario	Gaze Only	Casual Pointing	Precise Pointing	Desc. Term	Name only	No Desc Term
Open/ Known	.206	.706	.088	.402	.461	.137
Open/ Unknown	.265	.637	.098	.922	0	.078
Closed/ Known	.814	.167	.020	.245	.588	.167
Closed/ Unknown	.559	.373	.069	.951	0	.049
Object	.049	.333	.618	.980	0	.020

partially extended. Precise pointing was defined as when the speaker's arm and index finger were fully extended, based on Kendon's definition [1].

There was a range of variation in the amount of extension of the upper arm and the forearm among participants, so for simplicity, we categorized the pointing type as precise pointing only when the arm and the index finger were fully extended. All other pointing was coded as casual pointing.

D. Categorization of Descriptive Terms

We analyzed the video to identify whether people used a verbal descriptive term. Here, a "descriptive term" is defined as an utterance aside from the referent's name that uniquely singles out the referent from other people, e.g. based on relative location ("the person in front of the coffee shop") or a visible feature ("the person in the blue shirt").

If only the referent's name was used, it was classified as "name only". If the participant used only a general deictic reference term ("that person"), it was classified as "no descriptive term", since terms like "this" or "that" may not uniquely single out the referent among surrounding people [6].

E. Results and Analysis

For each of the 5 scenarios, a total of 102 reference behaviors were observed (6 questions for each of the 17 participants). Table 1 shows the relative frequencies of behaviors for each scenario (see Table 1). The most frequently used behaviors in each scenario are highlighted in red.

Object vs. person: Participants rarely used precise pointing when referring to people (precise pointing: <10% for all cases), compared with referring to objects (precise pointing: 61.8%). This suggests there is a social factor that causes the speaker not to want to point precisely, in which he might risk singling someone out.

Open vs. closed: In closed conversations, "gaze only" was most common, whereas in open conversations, "casual pointing" was most common. Our interpretation is that as pointing precision increases, the noticeability of the gesture also increases, hence increasing the likelihood of the referent becoming aware of the conversation. This suggests that in closed conversation, the speaker is more concerned about whether the referent becomes aware of the conversation than in open conversation.

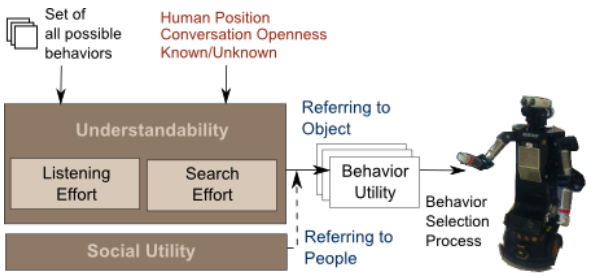


Fig. 3. Overview of Generative Model for Robot Behavior

Known vs. unknown: Interestingly, we did not see much difference in the use of gesture depending on whether the referent was known or unknown. However, the speaker used more descriptive terms when the referent was unknown to the listener than when the referent was known (e.g. for the Open/Unknown case, 92.2% used descriptive terms, while for the Open/Known case, only 40.2% used descriptive terms). Speakers still used pointing behavior even when using the referent's name (e.g. in the Open/Known case, casual pointing with name was used 32.4% of the time), even though the name would be enough to unambiguously identify the referent. Perhaps this was to make it easier for the listener to understand the reference, or to share the speaker's area of spatial attention.

IV. GENERATIVE MODEL FOR ROBOT BEHAVIOR

A. Overview

Previous studies have modeled pointing as a way to resolve ambiguity when referring to an object. We thus include *understandability* as the first factor in our model, which we define to encompass both resolution of ambiguity and ease of understanding. We then define an additional factor of *social utility*, which reflects the desire of the speaker to be polite by not singling the referent out (see Fig. 3). We believe that *social utility* is the main reason for the variations in deictic behavior between referring to people and referring to objects.

We propose a model to generate humanlike deictic behaviors in a robot by combining these factors of understandability and social utility into a behavior utility function. There is an inherent trade-off between these two factors. For example, pointing precisely at a particular individual may easily identify that person (high *understandability*), but the speaker may have made that person feel singled out and uncomfortable (low *social utility*).

To select a deictic behavior for a robot, the behavior utility function is evaluated for each of the potential deictic behaviors the robot can perform. We consider six behavior possibilities in our model: one of three pointing behaviors (gaze only, casual pointing, or precise pointing) combined with either the use or the non-use of a descriptive term.

B. Understandability

1) Overview

Regarding understandability, we generally assume that with some effort, the listener will eventually identify the target, but

pointing makes it easier to search for the referent since the listener can focus their search to a specific region that was pointed to. In this sense, pointing has reduced the listener's time and effort in searching for the referent. We introduce this concept of "search effort" as one element related to understandability. The more precisely the speaker points to the referent, the lower the listener's search effort will be.

The speaker's use of a descriptive term about the referent can also help reduce search time. We model this concept as "listening effort". Thus, we modeled the **understandability** as a function which decreases as the sum of these two effort factors. We assumed perfect understanding if no effort is required.

Understandability = 1– (Search Effort + Listening Effort) (1)

2) Search Effort

a) Modeling Based on Search Time

We modeled "search effort" based on the concept of a visual search task [14], in which an observer is searching for a target among a variable number of distractors (other people or features in the environment). Longer visual search times roughly equate to higher search effort. Hence, we approximate the **search effort** as proportional by a factor ω_1, with visual search time (t_{search}), as shown in Eq. 2

$$\text{Search Effort} = \omega_1 \times t_{search} \quad (2)$$

The variable number of distractors, or the **total amount of distraction** D_T, is the sum of both the number of human distractors and the environmental distraction. The **visual search time** for such a task is computed as the average reaction time, $t_{reaction}$, spent on each distraction, times the total amount of distraction (D_T). The modeling of $t_{reaction}$ will be explained in the following subsections.

$$t_{search} = t_{reaction} \times D_T \quad (3)$$

b) The Effect of Pointing Precision on Distraction

Pointing singles out a spatial area, but not necessarily a single entity in the world. Other studies have modeled pointing as a cone representing the angular resolution of the pointing gesture [15], which is centered along a beam originating from the pointing finger to the intended target, and has the angular width of a given resolution angle on either side of the beam. Previous findings indicate a resolution angle of a precise pointing cone of about 12 to 24 degrees [16]. We approximated the **pointing cone's resolution angle** $\theta_{pointing\ precison}$ to be 15 degrees to either side for precise pointing and 60 degrees to either side for casual pointing. For gaze only, we used an angle of 90 degrees, based on the human's forward-facing horizontal field of view.

Recall that our visual search time model is based on searching for a target among a number of distractions. Even when there is only one person in the environment, it will still take some time to find the referent, particularly when the speaker points casually to a referent located far away.

The **number of human distractors**, D_h, is defined as the number of people who could potentially be the referent and within the pointing cone's resolution angle $\theta_{pointing\ precision}$.

Since the environmental distraction is not discrete, we expect it to increase linearly with the pointing angular width. We model D_e, the **environmental distraction**, as a constant noise factor τ per unit angular resolution, integrated over the residual angular resolution of the pointing cone, excluding the angle $\theta_{referent}$ occupied by the referent (see Fig. 4 or 5 as examples), as shown in Eq. 4. The value of τ will be larger for more cluttered environments.

$$D_e = \tau(2 \cdot \theta_{pointing\ precision} - \theta_{referent}) \quad (4)$$

c) The Effect of Descriptive Term on Reaction Time

To distinguish the referent from other people, a speaker may use a unique description term in addition to pointing. Previous studies have shown that providing a cue [17] or being familiar with the target [18] can reduce the uncertainty of the target and consequently reduce the reaction time. If the referent is known to the listener, the speaker will use the referent's name to describe him in all cases (e.g. it will be unnatural to describe a mutual friend as "the man in blue shirt" rather than "Jack"). Thus, we model the **reaction time** $t_{reaction}$ to be shortest when the referent is known (see Eq. 5). When the referent is unknown to the listener, search time will be longer. However, use of a descriptive term will reduce $t_{reaction}$ compared with not using a descriptive term.

$$t_{reaction} = \begin{cases} t_k, & \text{if known + using name} \\ t_{ud}, & \text{if unknown + using descriptive term} \\ t_u, & \text{if unknown + no descriptive term} \end{cases} \quad (5)$$

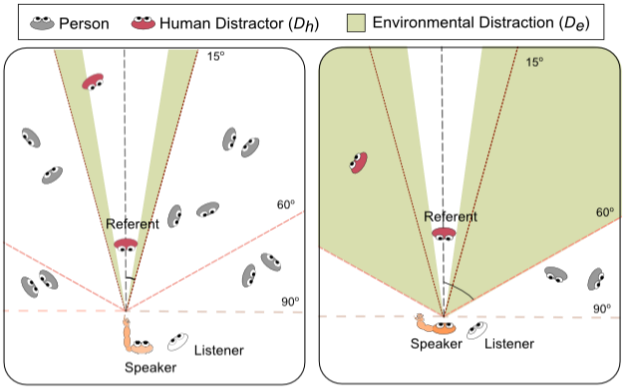
3) Listening Effort

The second factor in the *Understandability* equation is listening effort, representing the effort associated with the time required to listen to a descriptive term. For simplicity, we assign one of two discrete values to the **listening effort**: c_{desc} if a descriptive term is used, or $c_{no\ desc}$ otherwise in our model, as shown in Eq. 6. Since listening to a name or reference term requires less time, therefore less effort, than a descriptive term, we expect $c_{desc} > c_{no\ desc}$.

$$\text{Listening Effort} = \begin{cases} c_{\text{no desc}}, & \text{no descriptive term} \\ c_{\text{desc}}, & \text{using descriptive term} \end{cases} \quad (6)$$

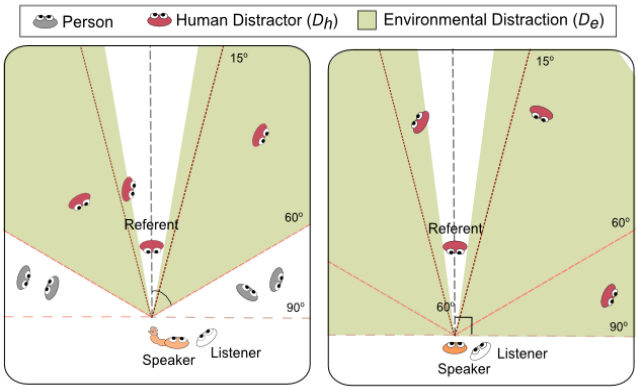
C. Social Utility

We model **social utility** as a quantity that will decrease if the speaker makes the referent feel uncomfortable or singled out. The loss in social utility is especially high in "closed" cases, when the content of closed conversation is leaked to the referent (e.g. the referent hears bad comments about him). To quantify this phenomenon, we consider the risk of the referent becoming aware of the conversation ($R_{awareness}$), multiplied by the cost to social utility (C_{social}) if the referent becomes aware, as shown in Eq. 7.



(a) Precise pointing is chosen (b) Casual Pointing is chosen

Fig. 4. Open/Unknown scenario: examples showing the influence of distractors on behavior selection



(a) Casual pointing is chosen (b) Gaze only is chosen

Fig. 5. Open/Known scenario: examples showing the influence of distractors on behavior selection

$$\text{Social Utility} = -(R_{awareness} \times C_{social}) \quad (7)$$

Recall that in our previous section we model precise pointing to have the effect of ruling out distraction. The presence of many distractors within the pointing cone, e.g. due to a less precise pointing gesture, makes it less clear whether the speaker is actually pointing to the referent, whereas a precise gesture with few distractors leaves little room for doubt. Thus we approximate the **awareness risk** ($R_{awareness}$) as the inverse of the total amount of distraction:

$$R_{awareness} = \frac{1}{(D_h + D_e)} \quad (8)$$

The **cost to social utility** is dependent upon the openness of the conversation. As explained above, the penalty to social utility due to the referent becoming aware of the conversation is much more severe in closed conversation than in open conversation. Thus, we model the cost to have one of two discrete values, based on the openness of the conversation, where $\beta_{closed} > \beta_{open}$.

$$C_{social} = \begin{cases} \beta_{closed}, & \text{if conversation is closed} \\ \beta_{open}, & \text{if conversation is open} \end{cases} \quad (9)$$

D. Calibration of Our Model

We calibrated our model based on the results of our data collection by choosing parameters for our model to maximize the correspondence between the most frequently predicted behaviors for each scenario (highlighted in red in Table 3) and the most frequently used behaviors in that scenario from the data collection (highlighted in red as shown in Table 1). Table 2 shows the calibrated parameters.

E. Examples of Using Our Model

The examples in Figure 4 and 5 illustrate situations where our model chooses different behaviors based on the amount of distraction and the scenario. The figure shows each person's position in the environment. The resolution angles for each of the three pointing cones (90° for gaze only, 60° for casual pointing, and 15° for precise pointing) are drawn as different shades of red dashed lines radiating out from the speaker.

Figure 4 shows examples in the Open/Unknown scenario. The most common behavior in this scenario is casual pointing. However, precise pointing is sometimes used in crowded environments, where it is harder to identify the referent. This is due to the distraction effect, as modeled previously.

Figure 4(a) is a case where the participant used precise pointing to identify the referent. In this crowded environment, there were 8 people within the region of casual pointing; thus, casual pointing would yield low understandability. However, precise pointing reduces the number of human distractors to 2, providing much higher understandability. Figure 4(b) illustrates a less crowded example. Here, due to the smaller number of distractors, the model chooses casual pointing, which yields enough understandability while yielding higher social utility.

Figure 5 shows two examples in the Open/Known scenario. As in the unknown scenario, the most common gesture is casual pointing. However, since the referent is already known to the listener, less ambiguity needs to be resolved. Figure 5(a) shows a crowded environment, but here casual pointing is enough to yield enough understandability. When the environment becomes less crowded, as in Fig. 5(b), using gaze only would be enough for understandability, while yielding high social utility.

TABLE II. CALIBRATED MODEL PARAMETERS

Search Effort		Social Utility		Listening Effort	
ω_1	.013	β_{open}	.273	c_{decr}	.011
t_k	.03	β_{closed}	30	$c_{no\ decr}$	0
t_{wd}	.07	w_{ref}	25[cm]		
t_u	.3				
τ	8.5				

V. EVALUATION WITH ROBOT

A. Hypotheses

In a field experiment, we compared the performance of our model against a model that considers only *understandability* but not *social utility*. This comparison model was chosen because it represents a typical state-of-the-art approach to generate deictic behaviors for referring to objects, and it will be referred to as the "object-reference model." We made the following hypotheses for the referent and listener:

TABLE III. RATIO OF PREDICTED BEHAVIORS FROM DATA COLLECTION USING CALIBRATED PARAMETERS

Scenario	Gaze only	Casual Pointing	Precise Pointing		Desc. Term	Name only	No Desc. Term
Open/ Known	.196	.804	0		0	1	0
Open/ Unknown	0	.804	.196		.99	.001	0
Closed/ Known	1	0	0		.001	.99	0
Closed/ Unknown	1	0	0		1	0	0
Object	0	0	1		.833	0	.167

Predictions for referent evaluations

- The referent will perceive the robot's behavior as *more polite*. Since the robot's pointing will be less precise, the referent is less likely to feel singled out.
- *Understandability* will be *lower* with the person-reference model, as the intention of social utility is to reduce the risk of the referent's awareness of conversation.
- The referent will perceive the robot's behavior to be *more natural* because the person-reference model is calibrated after observations of real human behavior.
- Politeness will be more important than understandability, since the referent is not directly involved in the conversation. Thus the referent will evaluate the proposed model as *better overall* than the object-reference model.

Predictions for listener evaluations

- Listeners will rate the robot as *more polite* with the person-reference model, due to sympathy with the referent, and because the listener will feel uncomfortable if information is leaked to the referent in closed conversations.
- *Understandability will be sufficient* with the person-reference model. Although there is a tradeoff between understandability and social utility, the model will provide enough understandability for the listener.
- The robot's behavior will be rated *more natural* because the person-reference model is calibrated after observations of real human behavior.
- As the person-reference model determines an appropriate balance between understandability and politeness, listeners will rate it *better overall* than the object-reference model.

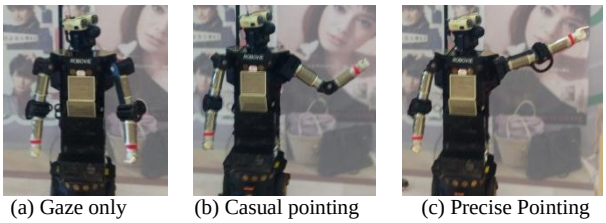
A. Experiment Setup

We implemented our model in a communication robot and hired participants to evaluate the robot's behavior in a series of short interactions. The experiment used a within-participants design and was counterbalanced between two conditions: *person-reference model* and *object-reference model*.

1) Implementation of Autonomous Robot System

For this experiment, we used Robovie 2, a humanoid robot with a 3-Degree-of-Freedom (DOF) head, two 4-DOF arms, a wheeled base, and a speaker that can output utterances. Robovie also has an actuator on its finger that allows it to do index-finger pointing. Six deictic behaviors were implemented into the robot, as categorized in our data collection, including three pointing behaviors: gaze only, casual pointing, and precise pointing (Figure 6), combined with the use or non-use of a descriptive term.

Our proposed model was implemented into the robot using all the equations with calibrated parameters. The robot autonomously executed the appropriate deictic behavior based on the output of the model.



(a) Gaze only (b) Casual pointing (c) Precise Pointing

Fig. 6. Examples of Robovie performing the three pointing behaviors

2) Procedure

We compared two conditions: the *person-reference-model* condition (our proposed model, including understandability and social utility) and the *object-reference model* condition (including understandability, but not social utility).

One participant acted as a listener and conducted short question-and-answer interactions with Robovie in a shopping mall. The other participant and a confederate acted as other customers. For each condition, Robovie and the listener asked each other a series of 8 questions: 2 questions each for four scenarios: Open/Known, Open/Unknown, Closed/Known, and Closed/Unknown, and each time Robovie made a reference to either the second participant or the confederate.

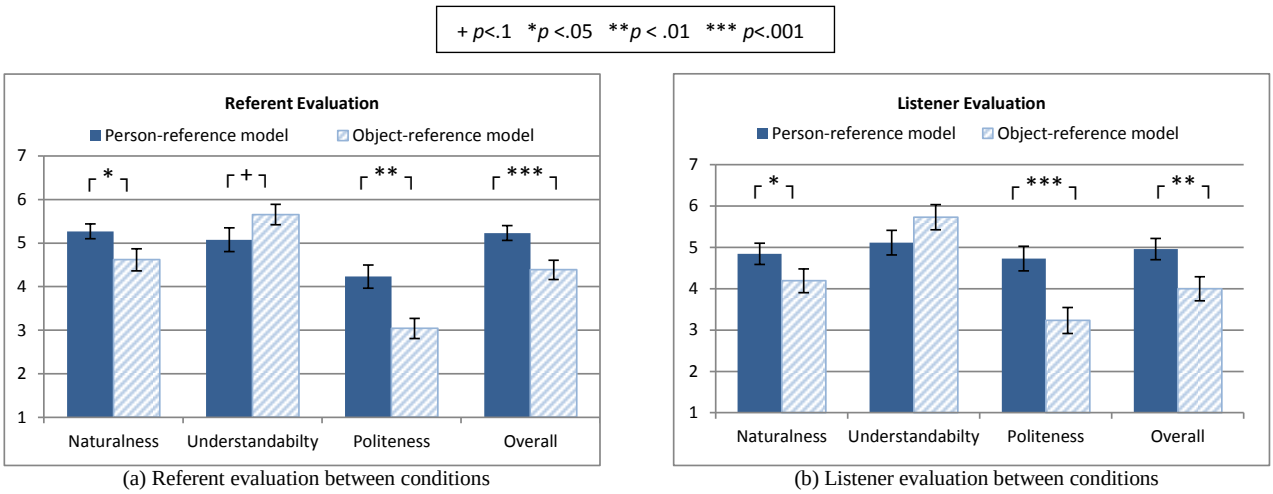
To prepare for the "known" scenarios, the participants and the confederate were asked to introduce themselves. This self-introduction was also intended to make the participants feel more investment in the conversation so they would become embarrassed if information were leaked in "closed" scenarios.

Participants' names were entered into the system before each trial, so the robot could refer to the referent by name in "known" scenarios. To standardize the descriptive terms for the "unknown" cases, the human distractors wore different colored badges so Robovie could refer to them by their badge color.

For "open" scenarios, the listener asked Robovie two pre-determined "neutral" questions. For the "closed" scenarios, Robovie asked the listener two pre-determined "sensitive" questions, e.g., "Which person do you think has bad fashion sense?" The listener answered by selecting either the second participant or the confederate. Because we believed that the listener might feel embarrassed by Robovie's impoliteness, Robovie then repeated the opinion stated by the listener while performing the selected deictic behavior, e.g. pointing while saying, "So you think Tanaka-san has no fashion sense?"

Since the volume of the robot's voice may affect evaluations, we adjusted the volume of the robot's voice to be louder in the "open" scenarios. For the "closed" scenarios, the volume was adjusted to a level that only the listener could hear.

After the four scenarios in one condition were completed, both participants answered questionnaires. The procedure was repeated with the remaining condition (*person-reference model*



(a) Referent evaluation between conditions (b) Listener evaluation between conditions

Fig. 7. Evaluation results of Robovie's behaviors between conditions

or *object-reference model*). The conditions were counter-balanced. At the end of the experiment, the participants were interviewed to gain a deeper understanding of their opinions.

With current speech recognition technology, it is difficult to accurately understand a person's speech in a noisy shopping mall. This noisy environment may risk the results of the experiments not making sense (e.g. if the robot misrecognized the name of the referent chosen by the listener).

To mitigate such risk, an operator assisted with speech recognition by listening to the listener's utterance transmitted through a teleoperation system. Upon hearing the listener's response for the chosen referent, the operator tagged the referent among the set of people detected by the human tracking system, and clicked "start" to trigger the execution of the robot's appropriate deictic behavior, which is determined autonomously by the proposed model.

3) Environment

All trials were conducted on weekdays in the same shopping mall location as the data collection. As the other people in the environment were shopping mall customers, we could not explicitly control the degree of crowding. However, we believe that the distribution of people in the environment was fair between conditions. On average, in the *person-reference model* condition, 6.61 people (s.d. 3.75) were present in the environment, compared with 6.53 people (s.d. 3.93) in the *object-reference model* condition.

4) Measurement

Both the listener and the referent rated the following items on a 1-7 scale (1 being very negative and 7 being positive for the respective items) in a written questionnaire:

- *Naturalness* of the robot's deictic behavior.
- *Understandability* of the robot's deictic behavior
- *Perceived politeness* of the robot's deictic behavior
- *Overall goodness* of the robot's deictic behavior

Because there were variations in the operator's speed and level of ambient noise, participants were asked not to consider timing or volume of the robot's utterances in their evaluations.

5) Participation

A total of 26 trials were conducted. 33 participants were hired (19 male, 14 female, average ages of 23 years old). 19 participants played the roles of listener and referent in different trials, but no participant played either role twice.

VI. Results

A. Verification of Hypothesis 1(Referent)

Figure 7(a) shows the questionnaire results from the referents. A one-way repeated-measures analysis of variance (ANOVA) was conducted with one within-participants factor, *model*, in two levels: *object-reference model* and *person-reference model*, for all measurements. The analysis revealed significant differences in *overall evaluation* ($F(1,25)=21.763$, $p<.001$, $\eta^2=.465$), *politeness* ($F(1,25)=15.391$, $p=.001$, $\eta^2=.381$), and *naturalness* ($F(1,25)=7.335$, $p=.012$, $\eta^2=.227$), and there was an almost-significant difference in *understandability* ($F(1,25)=3.362$, $p=.079$, $\eta^2=.119$).

These results support our hypothesis that the referents would perceive the overall behavior to be better with the person-reference model. The result also supports our predictions for *politeness* and *naturalness*, but not our prediction for *understandability*.

B. Verification of Hypothesis 2 (Listener)

Figure 7(b) shows the questionnaire results from the listeners. A one-way repeated-measures ANOVA was conducted for all measurements. There were significant differences in *overall evaluation* ($F(1,25)=10.192$, $p=.004$, $\eta^2=.290$), *politeness* ($F(1,25)=25.0$, $p<.001$, $\eta^2=.500$), and *naturalness* ($F(1,25)=4.972$, $p=.035$, $\eta^2=.166$), but no significant difference in *understandability* ($F(1,25)=2.235$, $p=.147$, $\eta^2=.082$).

These results support our prediction that listeners would rate the person-reference model better in *overall evaluation*, as well as our predictions for *politeness* and *naturalness*.

C. Discussion of Results

Many participants said that they rated our proposed model better because the robot behaved more politely. For listeners, it

was particularly embarrassing when the robot repeated his/her negative comment about the referent together with precise pointing. No significant difference was found for understandability. One possible reason is that the referents were asked to watch and evaluate the robot, so they were inevitably more aware of the conversation than a typical bystander would be.

D. Limitations and Future Work

In this study, the main focus was on generating the deictic behavior that best balances the issues of being polite and being easy to comprehend. This work could be extended to include more detail, such as exploring degrees of casual pointing or other deictic gestures such as chin-pointing. Our study also examined the effect of the use or non-use of descriptive terms, but future work could investigate relative effects of different kinds of descriptive terms or different levels of specificity.

One issue that we did not cover was the effect of gaze and how it relates to politeness and understandability. Gaze cues are a known attention drawing mechanism, and participants in our experiment explicitly noted that the robot's gaze helped them to identify the referents. In our experiment, we implemented the robot's gaze to look at the direction of the referent consistently in both conditions. However, we did not explore what type of roles gaze cues actually play in understandability and politeness.

Pointing behaviors carrying semantic meaning were not fully explored in our study. When a person introduces another person, they usually use an open hand, palm up gesture as an implication for offering. Including more social settings (e.g. introduction) would be an interesting area for future work.

While there are many possible areas for future work, we believe that our model is relatively accurate in representation of the main factors of real human deictic behavior.

VII. CONCLUSION

In this study, we have empirically confirmed that people's pointing behavior is different when they refer to objects and when they refer to people. From data of real human deictic behaviors, we developed a model for generating deictic behaviors for robot that best balance comprehension and politeness. We compared our model, which considers both understandability and social utility, with an object-reference model that aims to only maximize understandability. We demonstrated that with our model, the robot's behavior was perceived more polite and natural, and therefore the robot's behavior led to a better overall interaction.

ACKNOWLEDGMENT

We would like to thank Satoshi Koizumi for facilitating the smooth operation of the experiments. This work was supported by the Ministry of Internal Affairs and Communications of Japan.

REFERENCES

[1] A. Kendon, Gesture: Visible Action as Utterance. Cambridge: Cambridge University Press, 2004.

[2] I. van der Sluis and E. Krahmer, "Generating Referring Expressions in a Multimodal Context: An empirically motivated approach", In Proceedings of 11th CLI, 2001.

[3] S.L. Haywood, M.J. Pickering, and H.P Branigan, "Do speakers avoid ambiguities during dialogue?" Psychological Science, vol. 16, no.5 , pp. 362–366, 2005.

[4] I. Paraboni, K. van Deemter, and J. Masthoff, "Generating referring expressions: Making referents easy to identify", Computational Linguistics, vol. 33, no. 2, pp. 229–254, 2007.

[5] A. Bangerter, E. Chevalley, 'Pointing and describing in referential communication: When are pointing gestures used to communicate?', in Proceedings of the workshop on multimodal output generation, 2007

[6] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "Humanlike conversation with gestures and verbal cues based on a three-layer attention-drawing model", Connection Science, vol. 18, no. 4, pp. 379-402, 2006.

[7] J. Schmidt, N. Hofemann, A. Haasch, J. Fritsch, and G. Sagerer, "Interacting with a Mobile Robot: Evaluating Gestural Object References", IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2008), pp. 3804 - 3809.

[8] S. Sakurai, E. Sato, T. Yamaguchi, "Recognizing Pointing Behavior using Image Processing for Human-Robot Interaction", IEEE/ASME International Conference on Advanced Intelligent Mechatronics, pp. 1-6, 2007.

[9] T. Spexard, S. Li, B. Wrede, J. Fritsch, G. Sagerer, O. Booij, Z. Zivkovic, B. Terwijn, and B. Krose, BIRON, "Where are you? Enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization", IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2006), pp. 934-940, 2006.

[10] Y. Hato, S. Satake, T. Kanda, M. Imai, and N. Hagita, "Pointing to space: modeling of deictic interaction referring to regions," in Proc. of the 5th ACM/IEEE Intl. Conf. on Human-Robot Interaction, pp. 301–308, ACM, 2010

[11] A. Brooks, and C. Breazeal, "Working with robots and objects: Revisiting deictic reference for achieving spatial common ground", in Proc. of the 2006 ACM Conference on Human-Robot Interaction (HRI), 2006.

[12] A.C. Schultz, and J.G. Trafton, "Towards collaboration with robots in shared space: Spatial perspective and frames of reference". Interactions. Xii.2 (March-April), pp. 22-24, 2005.

[13] D. F. Glas, T. Miyashita, H. Ishiguro, and N. Hagita, "Laser-Based tracking of human position and orientation using parametric shape modeling", in *Advanced Robotics*, Vol. 23, No. 4, pp. 405-428, 2009.

[14] J.M Wolfe, "Guided Search 2.0. A revised model of visual search", Psychonomic Bulletin & Review, 1, pp. 202–238, 1994

[15] A. Kranstedt, A. Lucking, T. Pfeiffer, H. Rieser, and I. Wachsmuth. "Deixis: How to Determine Demonstrated Objects" Presented at Gesture Workshop 2005, Ile de Berder, France, 2005.

[16] P. Kuhnlein and J. Stegmann. "Empirical Issues in Deictic Gesture: Referring to Objects in Simple Identification Tasks". Technical Report 2003/3, SFB 360, University of Bielefeld, 2003.

[17] J. M. Wolfe, T. S. Horowitz, N. Kenner, M. Hyle, and N. Vasan, "How fast can you change your mind? The speed of topdown guidance in visual search". Vision Research, vol. 44, pp. 1411-1426, 2004.

[18] Q. Wang, P. Cavanagh, and M. Green, "Familiarity and pop-out in visual search", Perception Psychophys. *56*, pp. 495–500, 1994.