International Journal of Social Robotics (preprint)
The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-016-0348-9

A Model for Generating Socially-Appropriate Deictic Behaviors towards People

Phoebe Liu, Dylan F. Glas, Takayuki Kanda, *Member, IEEE,* Hiroshi Ishiguro, Norihiro Hagita, *Senior Member, IEEE*

ABSTRACT

Pointing behaviors are essential in enabling social robots to communicate about a particular object, person, or space. Yet, pointing to a person can be considered rude in many cultures, and as robots collaborate with humans in increasingly diverse environments, they will need to effectively refer to people in a socially-appropriate way. We confirmed in an empirical study that although people would point precisely to an object to indicate where it is, they were reluctant to do so when pointing to another person. We propose a model for selecting utterances and pointing behaviors towards people in terms of a balance between understandability and social appropriateness. Calibrating our proposed model based on empirical human behavior, we developed a system able to autonomously select among six deictic behaviors and execute them on a humanoid robot. We evaluated the system in an experiment in a shopping mall, and the results show that the robot's deictic behavior was perceived by both the listener and the referent as more polite, more natural, and better overall when using our model, as compared with a model considering understandability alone.

INTRODUCTION[1]

The importance of natural and humanlike human-robot interaction is gaining more attention as robots gain presence in museums [2-4], classrooms [5], and elderly care facilities [6, 7]. In order to facilitate natural and intuitive communication, humanlike spoken, locomotive [8], and gestural behaviors are being developed for robots, and one important area of focus is in deictic gestures, such as pointing. Several studies in human-robot interaction have focused on generating human-like multimodal referring acts using both speech and gesture for objects [9-13] and space [14, 15].

[1] This paper is an extended version of our conference paper [1] with integrated technical details, additional discussions, expanded explanations, and supplementary analysis of the experiment.

Our study focuses on a method for generating behaviors for a robot to point to a person. There are important differences in the way someone gestures towards objects and the way someone gestures towards a fellow person. When pointing to people, it is often considered more appropriate to gesture casually to them rather than using a very obvious pointing gesture, i.e. with an extended index finger. However, in most situations there would be no reason not to use a clear and precise pointing gesture when identifying an object.

As social human-robot interactions become more complex, it will be important to consider the social appropriateness of a pointing gesture within the context of the conversation. For example, if an elder-care provider is consulting with another practitioner about the health condition of a particular senior person, he would probably discreetly point out that person, using a subtle pointing gesture, in order to reduce the risk of the referent becoming aware and avoid causing anxiety to the referent. In such a scenario, if a robot directly singled out the individual when discussing a sensitive topic (i.e. a "closed" conversation), the robot would probably be perceived as socially-inappropriate. It would be more appropriate for the robot to discreetly identify the referent, even if it meant being less clear to its listener about the referent's identity. However, if the conversation was not of a sensitive nature, and the topic being discussed is neutral or positive (i.e. an "open" conversation), the social consequences would be less severe, and it might be acceptable for the robot to be more obvious about identifying the referent.

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.

First, we present an empirical study of human pointing behavior, in which we confirm that people usually do not use precise pointing gestures, that is, they typically do not use the 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 then propose a generative model for deictic behaviors, based on the idea of a balance between understandability and social appropriateness: more precise pointing gestures can increase understandability, but

they can also 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 describe our implementation of this model in a real robot system and 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.

RELATED WORK

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 [16]. 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 it provides a useful basis for our categorization.

Some studies have examined the use of 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 [17-19]. Accordingly, we also consider verbal descriptive terms as part of our model for generating deictic behavior.

Human-Robot Interaction

Various generative robot behaviors first look at how humans behave as the basis of behavior design. For example, Semel et al. developed and verified a control system for humanoid bipedal locomotion that was biologically based on human gait cycles [8]. However, the mechanism that drives us to act a certain way may not be obvious to us. Hence, various studies use data-driven methods to extract the underlying mechanisms that govern our behaviors, such as recognizing our emotional states through ECG data [20], or identifying features that uniquely define us through EEG data [21]. In our work, we first observe human deictic behaviors through data collection, and then we incorporate the main factors that were identified in our analysis into our model.

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 [22]. 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 focused on generating human-like multimodal referring acts using both speech and gesture for objects [9-12], and space [14, 15]. Brooks and Breazeal [23] 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 [24]. 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.

DATA COLLECTION

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 in which pointing behaviors vary when pointing to people. In particular, we were interested in examining three factors:

Object vs. person: As 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 wondered whether people's behavior would be different if they already knew the referent, such as in the case where saying their name would be enough to identify the referent without ambiguity.

Procedure

We conducted the data collection in a shopping mall, as shown in Fig. 1(a), with 17 participants (11 female, 6 male, average 23.7 years old), who were paid. We asked the participants to role-play as customers in the shopping mall. An experimenter asked the participant's opinions about other products or visitors in the mall, and the participant freely answered using deictic behaviors. The participants were not explicitly instructed to use deictic behaviors, but rather instructed to "indicate" who the referent was.

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 acquaintances) 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 acquaintances) 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 pre-determined questions, which were counter-balanced. Before the experiment, we had a short ice-breaker session to familiarize the participant with two additional experimenters, who were role-playing as the acquaintances in the "known" scenarios. The two acquaintances stood at different locations for each question. In the "unknown" scenarios, the participants were instructed to refer to actual customers in the shopping mall. 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 would be more difficult than when only a few customers were present), we used a human tracking system based on 2D laser range finders (LRF) [25] to capture the positions of the people in the environment. Fig. 1 (b) shows the map of the environment in which the data collection was conducted.

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.

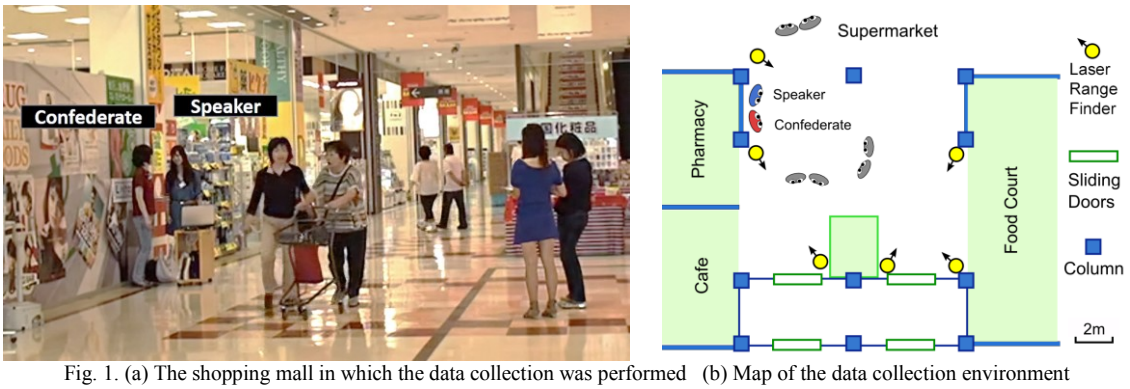


Fig. 1. (a) The shopping mall in which the data collection was performed (b) Map of the data collection environment

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.

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 partially extended. These also corresponded with the "Open Hand Neutral", "Open Hand Prone", and "Open Hand Oblique" pointing gestures as defined by Kendon.[16] "Precise pointing" was defined as when the speaker's arm and index finger were fully extended, based on Kendon's definition.

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.

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].

Fig. 2. Categorization of different pointing types

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). Using the recorded videos, an experimenter annotated the pointing behaviors and whether descriptive terms were used by the participants in each trial. This was used for the tabulation of Table I. The experimenter also noted down the referent's position at the time when the speaker made the reference behavior, as well as how long it took for the speaker to make the reference behavior. We noticed that in addition to the use of deictic pointing behaviors to describe the referent, some speakers also used other techniques of representation, such as using gesture to act out putting on a jacket to describe a referent wearing a jacket. These types of gestures were only observed a few times among the participants, and were not a universal phenomenon. In this paper, we avoid these special cases and focus only on deictic language and referential gestures.

The relative frequencies of behaviors for each scenario are shown in Table I, with the most frequently used behaviors in each scenario highlighted in bold.

Object vs. person: Participants rarely used precise pointing when referring to people (precise pointing: <10.0% 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.

In the closed scenario, we also observed that the speaker would often lean closer to the confederate when trying to identify the referent. This phenomenon was not observed in the open scenario. This was more evident when the referent was nearby in closed conversations. Studies have indicated that the forward body lean conveys a sense of intimacy, attraction, and trust [26, 27]. Due to the sensitive information that was being exchanged in the "closed" conversation, we speculate that the participants exhibited such behaviors due to feeling a greater sense of trust or affiliation with the confederate.

Interestingly, in closed conversation, some speakers would also giggle or nervously laugh when they were describing someone negatively (*e.g.* "I think that person with the shopping cart has no fashion sense at all."). We did not observe speakers laughing or giggling nervously in the open conversation, suggesting that the speakers had higher level of discomfort when describing the referent in the closed conversation than the open conversation [28].

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).

In general, we found that the speaker took more time to identify an unknown referent. When the referent was unknown to the confederate, the speaker would often repeat or elaborate on describing the referent. For example, the speaker saying, "the person wearing the blue jacket is the person I saw on the bus today," would sometimes be followed by the confederate confirming, "you mean that person in blue?" The speaker would then describe the referent in further detail such as, "he is also wearing glasses." On average, the speaker spent 6.25 seconds describing an unknown referent, and 4.41 seconds describing a known referent. Some 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.

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

GENERATIVE MODEL FOR ROBOT BEHAVIOR

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. For example, a crowded environment where a lot of effort is required to identify a person will lower the ease of understanding for the listener.

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.

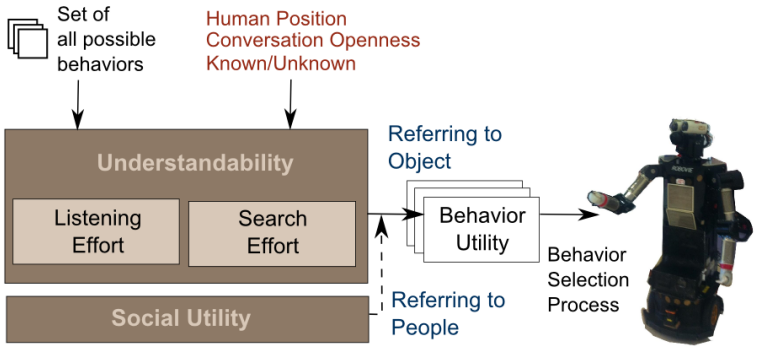


Fig. 3. Overview of Generative Model for Robot Behavior

Understandability

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. The speaker's use of a descriptive term about the referent can also help the listener reduce search effort, since providing cues can help to quickly distinguish the referent among other people or objects. We introduce this concept of "search effort" as one component of understandability. The more search effort is required, the less understandability the listener will have.

Although the use of a descriptive term may help decrease search effort, it also imparts extra cognitive load on the listener to interpret the descriptive term, and hence decreases their ease of understanding. We designate this component of understandability as "listening effort". 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.

$$Understandabilty = 1 - (Search\ Effort + Listening\ Effort) \quad (1)$$

Eq. 1 does not include explicit weighting factors for these two terms because, as we will explain below, our definitions of Search Effort and Listening Effort implicitly include parameters which can be tuned to adjust their relative weights in contributing to understandability.

Search Effort

Modeling Based on Search Time

We modeled "search effort" based on the concept of a visual search task [29], 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 linearly proportional by a factor w_1, with visual search time (t_{search}), as shown in Eq. 2. w_1 is a parameter which will be tuned.

$$Search\ Effort = w_1 \times t_{search} \quad (2)$$

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. To search for a target among distractions, the listener spends attention and time, $t_{reaction}$, from item to item until the target is found or until all items have been checked [30, 31]. 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), as shown in Eq. 3. The modeling of $t_{reaction}$ will be explained in the following subsections.

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

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 [32], 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 [33]. We approximated the pointing cone's resolution angle $\theta_{pointing\ precision}$ 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, D_T. 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 $\theta_{\text{pointing precision}}$.

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, as shown in Eq. 4. The value of τ will be larger for more cluttered environments.

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

Recall in the previous section that the total amount of distraction D_T is the sum of both the number of human distractors, D_h and the environmental distraction, D_e. Thus, D_T will be directly influenced by the pointing gesture used by the speaker. An example from our data collection, shown in Fig. 4, illustrates how D_T is affected by the different sizes of the pointing cones. In this example, 6 people are present in the speaker's forward horizontal field-of-view of 180 degrees in our shopping mall environment. Using gaze only, all 6 people within

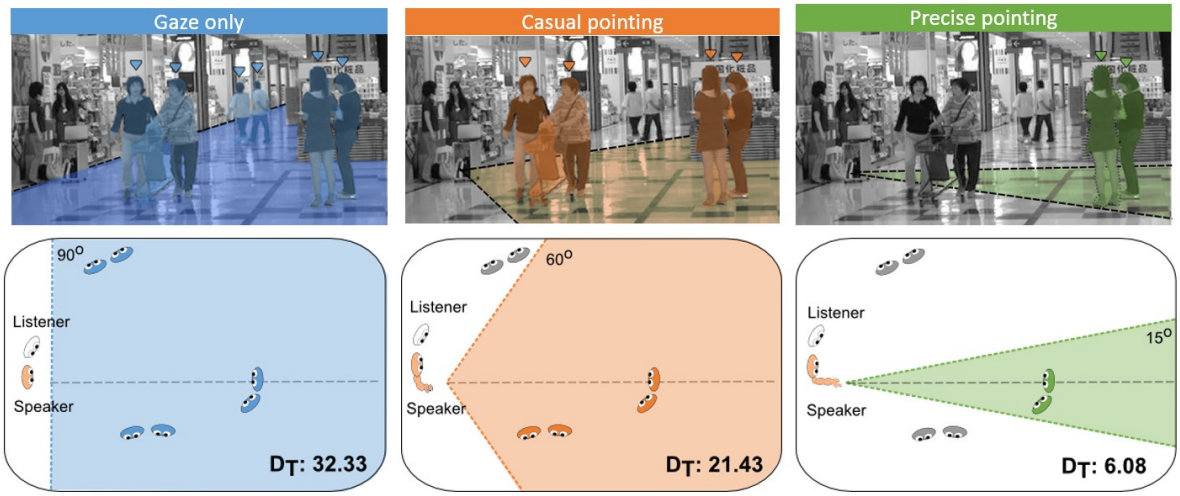


Fig. 4. An example of D_T for each pointing type in an environment with a total of 6 people. The highlighted people fall within the resolution angle of the pointing cone and are used for calculating D_T. Using "gaze only" leads to the highest D_T.

the speaker's view will be included as human distractors, whereas casual pointing reduces D_h to 4 people, and precise pointing reduces D_h to 2 people. Likewise, D_e is affected by the pointing type according to equation (4), in this case, 26.33 for gaze only, 17.43 for casual pointing, and 4.08 for precise pointing.

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 [34] or being familiar with the target [35] 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** $\mathrm{t}_{\mathrm{reaction}}$ 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 $\mathrm{t}_{\mathrm{reaction}}$ compared with not using a descriptive term.

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

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 $c_{desc} > c_{no\ desc}$.

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

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.

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

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** ($R_{\text{awareness}}$) as the inverse of the total amount of distraction:

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

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 $\beta_{\text{closed}} > \beta_{\text{open}}$.

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

Calibration of Our Model

We manually calibrated our model based on the results of our data collection by adjusting parameters for our model until the correspondence between the most frequently predicted behaviors for each scenario (highlighted in bold in Table II) and the most frequently used behaviors in that scenario from the data collection (highlighted in bold in Table I) were maximized. Table III shows the calibrated parameters.

TABLE II. 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

TABLE III. CALIBRATED MODEL PARAMETERS

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

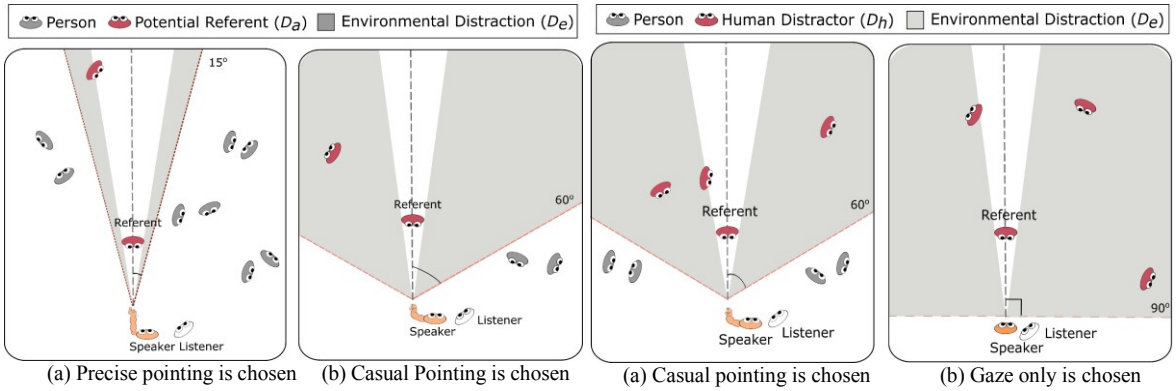
Examples of Using Our Model

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 (90^0 for gaze only, 60^0 for casual pointing, and 15^0 for precise pointing) are drawn as different shades of red dashed lines radiating out from the speaker.

Figure 5 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.

Fig. 5(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. Fig. 5(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.

Fig. 6 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. Fig. 6(a) shows a crowded environment, but here casual pointing is enough to yield enough understandability. When the environment becomes less crowded, as in Fig. 6(b), using gaze only would be enough for understandability, while yielding high social utility.



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

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

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

Fig. 6. Open/Known scenario: examples showing the influence of distractors on behavior selection

Model validation

The goal of our model is to generate a reasonable policy for producing socially-appropriate behaviors, rather than exactly replicating individual people's deictic behaviors. It is often difficult for a system to replicate exactly what humans do due to natural variation or randomness that arises among individuals. For instance, in the "Open/Unknown" scenario, there were 5 trials where 5 human distractors were tracked in the environment. Of the 5 trials, 1 participant used "gaze only", 3 participants used "casual pointing", and 1 participant used "precise pointing". This suggests that some deictic behaviors may be used interchangeably in some situations or dependent on the personality or culture of individuals. For this reason, we aimed to generate robot behaviors based on the dominant behavior trends observed from the data collection.

Table IV shows the confusion matrix of the predicted behavior using our model, based on the observed behavior from our data collection. The overall prediction accuracy was 81.3% for the "Closed/Known" scenario, 55.8% for the "Closed/Unknown" scenario, 42.1% for the "Open/Known" scenario, and 52.0% for the "Open/Unknown" scenario.

As a result of our calibrated parameters, our model tends to perform on the side of caution (*i.e.* the robot chooses deictic gestures that are less socially awkward). In both "Closed/Known" and "Closed/Unknown" scenarios, our model always selects "gaze only." This is consistent with the human behaviors observed in the data collection, where "gaze only" is the most frequently observed human behavior. Furthermore, in the data collection, people avoided using precise pointing for both "Closed" scenarios, and our model also behaves in the same way - the specificity (true negative rate) for precise pointing in "Closed/Known" was 98.0% and for "Closed/Unknown" was 93.1%.

In the "Open" scenarios, casual pointing constituted the majority of observed deictic behaviors (70.6% for "Open/Known" and 63.7% for "Open/Unknown"), and our model similarly predicted casual pointing the majority of the time (80.4% in both scenarios). The model was less successful in reproducing the other pointing behaviors, and we believe this variability could be due to individual preferences, or possibly related to unmodeled factors such as the precision of the descriptive terms used. It is also possible that gaze only and casual pointing can be used interchangeably in some situations, in which case multiple behaviors might be socially appropriate.

TABLE IV. CONFUSION MATRIX FOR OBSERVED BEHAVIOR FROM DATA COLLECTION AND MODEL PREDICTION

Closed/Known			
Data Collection / Model Prediction	Gaze only	Casual pointing	Precise pointing
Gaze only	83	17	2
Casual pointing	0	0	0
Precise pointing	0	0	0

Open/Known			
Data Collection / Model Prediction.	Gaze only	Casual pointing	Precise pointing
Gaze only	8	11	1
Casual pointing	13	61	8
Precise pointing	0	0	0

Closed/Unknown			
Data Collection / Model Prediction	Gaze only	Casual pointing	Precise pointing
Gaze only	57	38	7
Casual pointing	0	0	0
Precise pointing	0	0	0

Open/Unknown			
Data Collection / Model Prediction	Gaze only	Casual pointing	Precise pointing
Gaze only	0	0	0
Casual pointing	21	52	9
Precise pointing	6	13	1

SYSTEM ELEMENTS

Fig. 7 illustrates the system architecture for autonomously generating the robot's pointing behavior and utterances. We set up the **human tracking system** in the entrance hall of a shopping mall, covering an area of approximately 15m by 15m, as shown in Fig. 1. Pedestrian tracking was performed using the ATRacker[2] human tracking system presented in [25, 36], utilizing 6 laser range finders (LRF's) mounted in portable poles placed around the environment. This system combines range data from multiple sensors to track the trajectories of potential distractors in the environment using particle filters, and can provide position data within 6 cm error at a data rate of 37 Hz.

A **dialogue generator** able to produce utterances for the robot to speak was also implemented in the robot platform. This was used for producing questions to start each trial, such as, "who did you see at the bus stop yesterday?", as well as for generating deictic utterances based on the openness of the conversation and the familiarity of the referent. When necessary, the content for descriptive terms was automatically generated based on information entered before the experiment by a human experimenter (i.e., the person's name and their badge color).

[2] ATRacker is a product of ATR Promotions: http://www.atr-p.com/products/HumanTracker.html

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, a human operator acts as a **speech recognizer** by listening to the listener's utterance transmitted through a GUI. Upon hearing the listener's response for the chosen referent, the operator manually tags the referent among the set of people detected by the human tracking system, and clicks "start" to trigger the calculation of the most appropriate deictic behavior in the **generative model**, which was implemented in the robot using all the equations with calibrated parameters. Through its **speech synthesizer** and **actuator**, the robot autonomously executes the selected deictic behavior based on the output of the model.

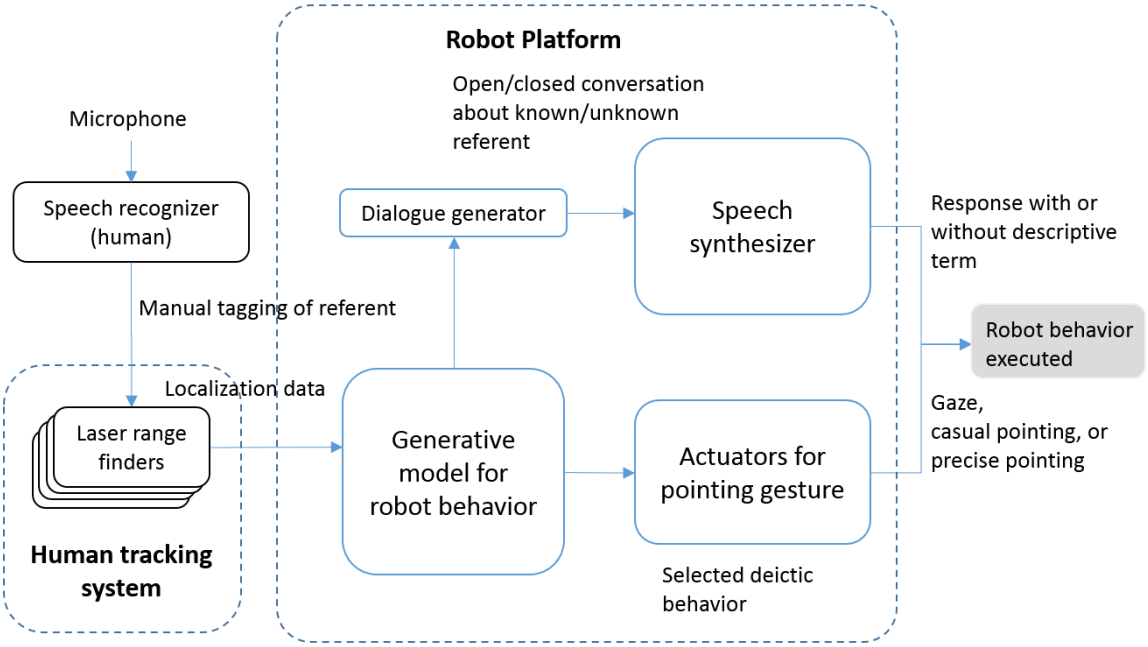
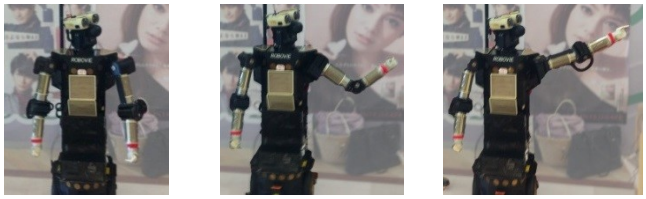


Fig. 7. System architecture for person-reference model: Inputs from speech recognizer and human tracking system are fed into the generative model, which then automatically calculates the appropriate deictic behaviors. The robot then responds verbally through its speech synthesizer and generates gaze and pointing gestures with its actuators.

Robot Platform

The robot platform we used was Robovie 2, a humanoid robot with a 3-Degree-of-Freedom (DOF) head, two 4-DOF arms, a wheeled base, and a speech synthesizer. We implemented motion behaviors for the three pointing behaviors: gaze only, casual pointing, and precise pointing (Fig. 8), and we implemented utterance behaviors incorporating the use or non-use of a descriptive term. Robovie's pointing gestures were implemented to best resemble what was commonly observed among the human participants.



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

Fig. 8. Examples of Robovie performing the three pointing behaviors

EVALUATION WITH A ROBOT

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:

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.

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*.

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 invested 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 poor 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* 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.

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.

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.

Participation

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

RESULTS

Verification of Hypothesis 1(Referent)

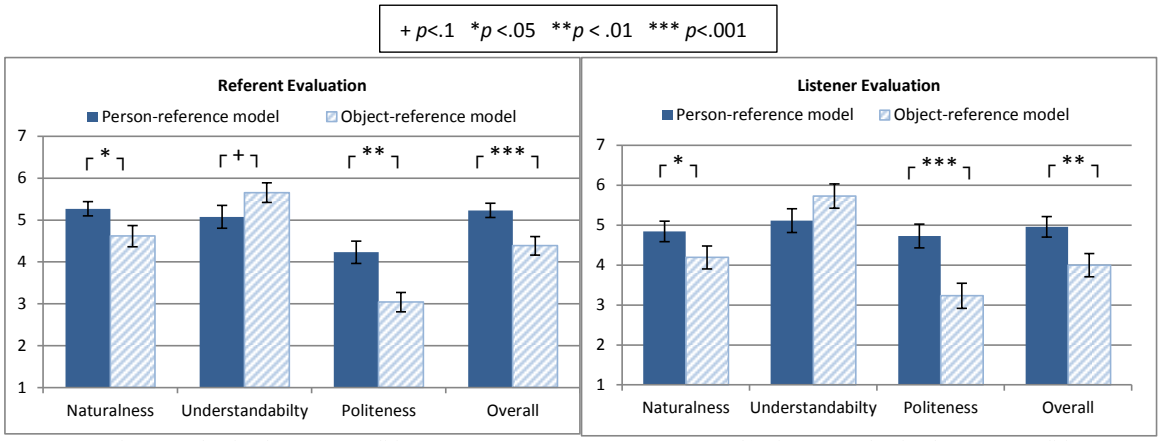
Figure 9(a) shows the questionnaire results from the referents. A one-way repeated-measures analysis of variance (ANOVA) was conducted with one within-participants factor, *model*, in two levels: *object-reference model* and *person-reference model*, for all measurements. The analysis revealed significant 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*.

Verification of Hypothesis 2 (Listener)

Figure 9(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*.



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

Fig. 9. Evaluation results of Robovie's behaviors between conditions

Analysis of understandability and social utility on the behavior-selection process

Our comparison experiment demonstrates how our *person-reference model* that considers both understandability and social utility can be used to improve the overall robot's deictic behavior, as compared with a model considering understandability alone. We provide a numerical analysis on the interactions observed in our experiment, in order to demonstrate how the values of understandability and social utility contributes to the behavior-selection process of our *person-reference model* under different scenarios.

To illustrate the tradeoff between understandability and social utility, Fig. 10 shows plots of the numerical values of understandability, social utility, and total behavior utility as a function of the amount of distraction in the environment (based on a "gaze only" pointing cone) for all of the "Open/Known" trials in our experiment. In this scenario, the robot's verbal behavior defaults to using the referent's name (i.e. without the use of a descriptive term), thus only 3 deictic behaviors are possible. Based on equations (1) – (3), we expect understandability to decrease linearly with D_T, as observed in Fig. 10(a). Precise pointing, with the smallest $\theta_{\text{pointing precision}}$, results in the highest value in understandability, followed by casual pointing and gaze only. From equations (7) and (8), we expect the negative effect of social utility to become weaker as D_T increases, as seen in Fig. 10(b), since the greater amount of crowding reduces the feeling of being singled out. Note that in Fig. 10(b), many of the data points for precise pointing fall below the bottom of the graph, since precise pointing leads to the lowest social utility.

The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-016-0348-9

Finally, our *person-reference model* considers both understandability and social utility. As shown in Fig. 10(c), the behavior utility of "gaze only" decreases as the environment becomes more crowded, whereas the behavior utility of "casual pointing" has an increasing trend as D_T increases. The model selects the behavior with the highest behavior utility, resulting in "gaze only" when D_T is low, and "casual pointing" as D_T increases. The behavior utility of "precise pointing" is too low to be shown on the figure, although it might be selected in extremely crowded situations. By contrast, the *object-reference model* would choose "precise pointing" in all cases, to maximize understandability regardless of social factors.

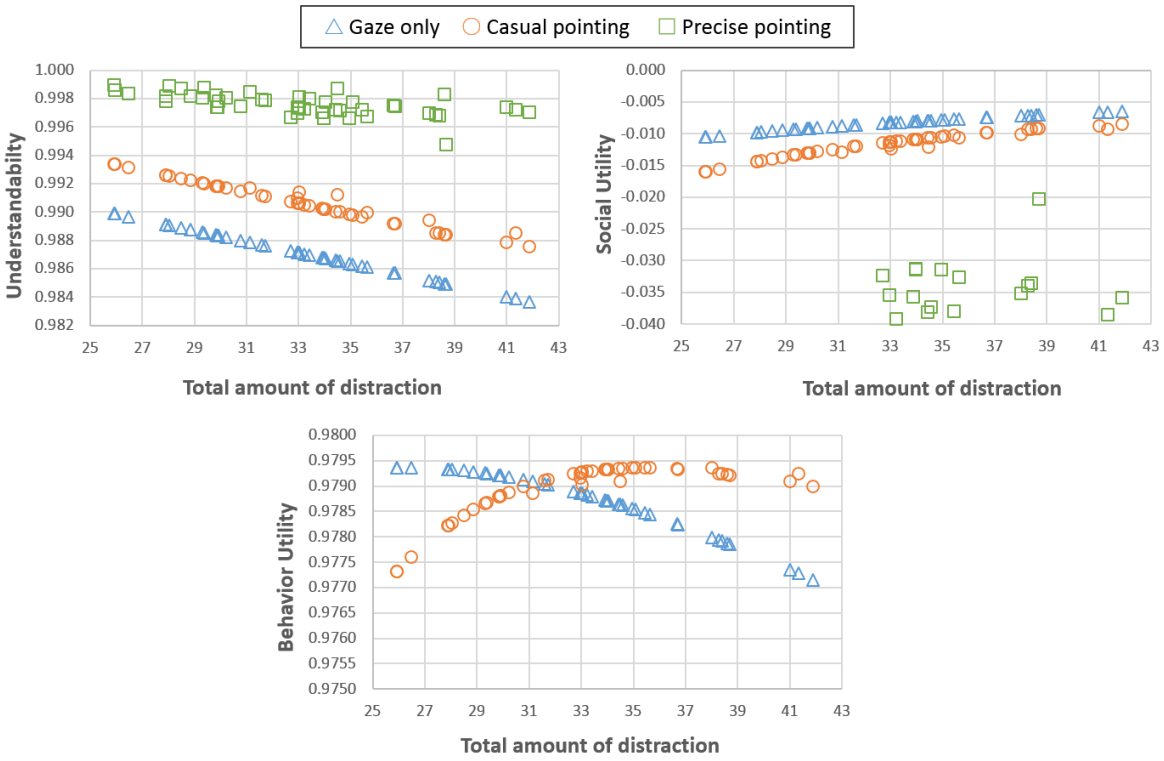


Fig. 10. The x-axis is the total amount of distraction, D_T, in the environment observed by the robot's forward facing horizontal field-of-view of 180 degrees at behavior execution time. For the three pointing behaviors in "Open/Known" scenario, the values of: (a) "Understandability" is negatively linear proportional to D_T, (b) "Social Utility" is inversely proportional to D_T, (c) "Behavior Utility", gaze only has the highest behavior utility when D_T is low, and casual pointing has the highest behavior utility as D_T increases. "Precise pointing" is too low to be shown on this figure, although it might be selected in extremely crowded situations.

Next, we provide an analysis of the behavior-selection process of our *person-reference model*, or the behavior utility values, for all scenarios. When the referent is unknown, a verbal description can also be used to resolve ambiguity of the referent. Verbal description increases understandability, but has no effect on social utility.

Therefore, the robot will use verbal description along with a pointing gesture for all unknown referents, as can be seen in Fig. 11 (c).

In "closed" conversation, the robot always chooses to use "gaze only," as shown in Fig. 11 (a) and (c). The high value of C_{social} leads to a greater influence of social utility than understandability in the values of behavior utility. As a result, the behavior utility follows the same trend of its social utility.

In "open" conversation, a tradeoff between understandability and social utility is observed in the behavior-selection process. In the "Open/Known" case, the robot uses "gaze only" when D_T is low. As D_T increases and requires a more precise pointing gesture to resolve ambiguity, the robot uses casual pointing (Fig. 11 (b)). In the

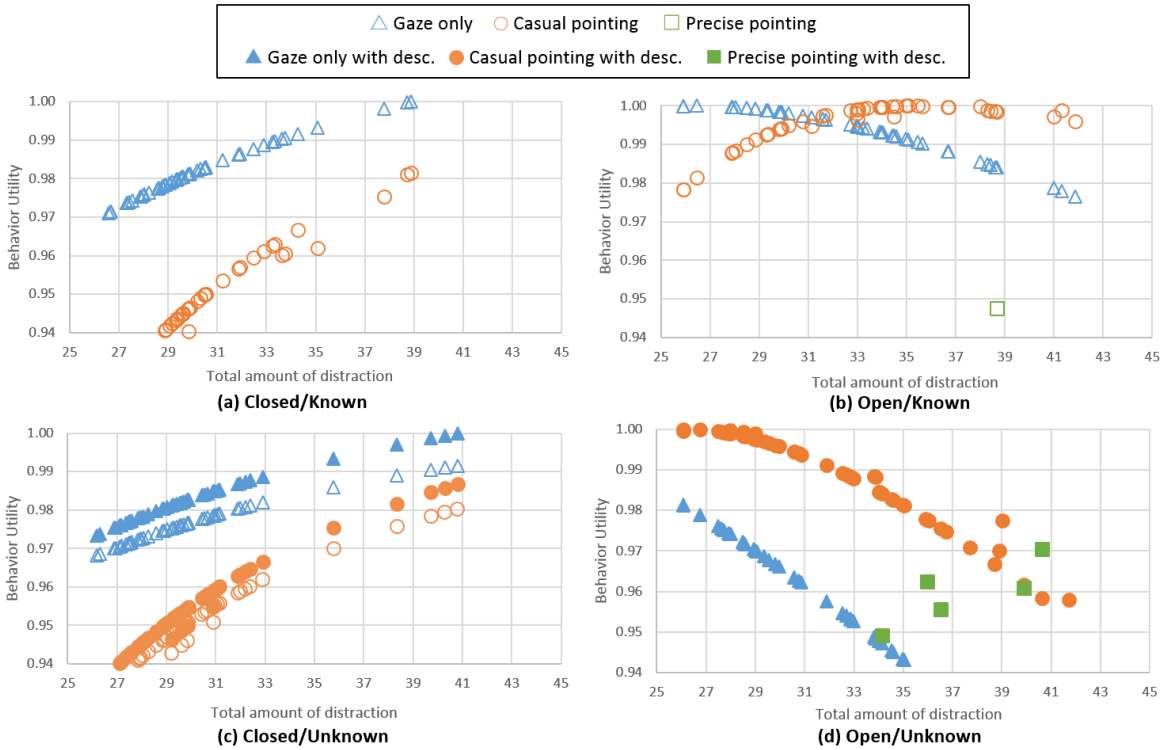


Fig. 11. Behavior Utility (normalized by scenario) of the three pointing behaviors calculated by the model for all experimental trials. The deictic behavior with the highest behavior utility is chosen by the robot. (a) "Closed/Known": gaze is always selected, (b) "Open/Known": Gaze is chosen when D_T is low, but casual pointing is chosen as D_T increases, (c) "Closed/Unknown": Gaze-only with description is always chosen, (d) "Open/Unknown": Casual pointing is mostly chosen, but when the total amount of distraction is high, precise pointing with description is chosen.

"Open/Unknown" case, the robot mostly used casual pointing together with a verbal description in our experimental environment. However, an increasing trend of using precise pointing together with a verbal description can be observed when there is a large amount of distraction (Fig. 11(d)).

DISCUSSION

Interview results from the participants

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. One participant commented she was worried the referent might get angry if he overheard the negative comments about him. It is interesting to note that one participant perceived the robot to be more "child-like" in the *object-reference model*, since the participant associated impoliteness in the robot's pointing behavior with the behavior of a child.

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.

Politeness in Pointing

When we first tried to set up a preliminary observation of people's pointing behavior, we set up a scenario where we asked the participants to imagine role-playing as a store clerk who was trying to indicate a store manager to a customer. In this scenario, we found that often the participants role-playing as the clerk were reluctant to use the index-finger pointing to identify the store manager in almost all scenarios, even when it was ambiguous who the store manager was due to the crowds. Instead, the participants used the more polite form of pointing, often with their palm up and hand open, to show the customer the whereabouts of the store manager. This was categorized as "Open Hand Supine" by Kendon, which may be semantically described as presenting or being ready to receive [16, 37]. It is possible that the participants were using this gesture to present the manager (i.e. the referent) to the customer.

Keeping in mind of these observations, it is important to consider the role and purpose of the robot when designing social behaviors. As described, a person may use more polite pointing gestures when interacting with a professional relation like a superior or customer, as compared to interacting with a friend. Thus, for example, a customer service robot might need to give priority more polite pointing behaviors, whereas a personal companion or non-humanoid robot might choose to maximize the understandability in its pointing behavior.

Limitations and future work

In this study, we developed a model for choosing deictic gesture and utterance behaviors that balance the issues of being polite and being easy to comprehend. While our study used a general categorization of gestures such as casual and precise pointing, there are many details which could be investigated in future work regarding the implementation and details of those gestures. For example, Kendon contrasted the semantic implications for different orientations of open-hand pointing – when a person introduces another person, they usually use an open hand, palm up gesture as an implication for offering [16], whereas when a person makes a critical remark, they use an oblique open-handed pointing. Incorporating the semantic meanings of the pointing gestures into our models may extend the robot's role and the scope of its interaction. For instance, a robot shop assistant presenting the manger or an educator making a critical remark may require the robot to adapt different subtle pointing hand orientations. Our study also examined the effect of the use or non-use of descriptive terms, but future research could investigate the relative effects of different kinds of descriptive terms or different levels of specificity. It may also be possible for models to be developed to quantify the degree of precision of a given pointing gesture, enabling more precise estimation of the pointing cone.

We understand that the use of kinesics or deictic behaviors may vary among cultures. The participants of this study were all Japanese, in which using body language may be remarkably restrained away from their in-group [38]. It is also worth noting that Japanese people may refrain from making hand gestures when the third-person referent is present, possibly to reduce the opportunity for offending anyone present and help sustain contextual harmony [39]. Imaginably, if this study was conducted in another culture, we might observe participants using more precise pointing. However, we believe our model does represent a universal phenomenon of how people point toward others, and hence, a valid model for robot deictic behavior as well.

In our experiment, a human operator was employed for two tasks: (1) to input the participant's name and the color of their name tag into the system before each trial, for later use in generating descriptive terms; and (2) to act as a speech recognizer in real time and tell the robot which person the chosen referent was. If this technique were to be used in a real social robot application, we expect that the referent's name would already be known, and the other functions could potentially be automated, e.g. using computer vision for identifying clothing color, and using gesture recognition and speech recognition to understand who the referent is. While implementing

these functions robustly is not trivial, we expect that with improvements in sensor technology the technique could be employed in an autonomous way.

While there are several possible directions for future work and refinement of the techniques presented here, we believe that this study has provided a successful demonstration of a practical technique for reproducing an important phenomenon which occurs in real human deictic behavior.

CONCLUSION

In this work, we have presented a model enabling robots to generate socially-appropriate deictic behaviors for referring to people, based on the openness of the conversation and familiarity with the referent, as well as the positions of people in the environment. In an empirical data collection, we observed that people's behavior varied both in terms of their pointing behaviors ("gaze only", "casual pointing", and "precise pointing") and their use of descriptive terms. We confirmed that people's deictic behaviors towards another person differed from their deictic behaviors towards objects, and we observed variation according to social context and presence of other people in the environment. From this data we developed a model enabling a robot to select socially-appropriate deictic behaviors towards humans based on a balance between understandability and social appropriateness for a given scenario.

Finally, we evaluated our model using a real robot in a shopping mall in an experimental comparison between our proposed model and a simpler model based only on understandability. The results showed significant differences for perception of the robot's deictic behaviors, in which the robot's behaviors were perceived to be more natural ($p < 0.05$ for both the referent and listener), polite ($p < 0.01$ for the referent and $p < 0.001$ for the listener) , and better overall ($p < 0.001$ for the referent and $p < 0.01$ for the listener) when using our proposed model. These results confirm that by considering social appropriateness in the model we were able to generate better social behavior for the robot.

ACKNOWLEDGMENT

We would like to thank Satoshi Koizumi for facilitating the smooth operation of the experiments. This study was funded in part by the Ministry of Internal Affairs and Communications of Japan and in part by JSPS KAKENHI Grant Number 25240042.

COMPLIANCE WITH ETHICAL STANDARDS

Conflict of Interest: The authors declare that they have no conflicts of interest.

This research was conducted in compliance with the standards and regulations of our company's ethical review board, which requires every experiment we conduct to be subject to a review and approval procedure according to strict ethical guidelines.

REFERENCES

[1] P. Liu, D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "It's not polite to point: generating socially-appropriate deictic behaviors towards people," in *8th ACM/IEEE International Conference on Human-Robot Interaction*, 2013, pp. 267-274.

[2] M. Bennewitz, F. Faber, D. Joho, M. Schreiber, and S. Behnke, "Towards a humanoid museum guide robot that interacts with multiple persons," in *Humanoid Robots, 2005 5th IEEE-RAS International Conference on*, 2005, pp. 418-423.

[3] M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita, "Interactive Humanoid Robots for a Science Museum," *Intelligent Systems, IEEE,* vol. 22, pp. 25-32, 2007.

[4] M. Nieuwenhuisen and S. Behnke, "Human-like interaction skills for the mobile communication robot robotinho," *International Journal of Social Robotics,* vol. 5, pp. 549-561, 2013.

[5] T. Kanda, R. Sato, N. Saiwaki, and H. Ishiguro, "A Two-Month Field Trial in an Elementary School for Long-Term Human-Robot Interaction," *Robotics, IEEE Transactions on,* vol. 23, pp. 962-971, 2007.

[6] K. Berns and S. A. Mehdi, "Use of an Autonomous Mobile Robot for Elderly Care," in *Advanced Technologies for Enhancing Quality of Life (AT-EQUAL), 2010*, 2010, pp. 121-126.

[7] A. M. Sabelli, T. Kanda, and N. Hagita, "A conversational robot in an elderly care center: An ethnographic study," in *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*, 2011, pp. 37-44.

[8] V. B. Semwal, S. A. Katiyar, R. Chakraborty, and G. Nandi, "Biologically-inspired push recovery capable bipedal locomotion modeling through hybrid automata," *Robotics and Autonomous Systems,* vol. 70, pp. 181-190, 2015.

[9] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, N. Hagita, and Y. Anzai, "Humanlike conversation with gestures and verbal cues based on a three-layer attention-drawing model," *Connection Science,* vol. 18, pp. 379-402, 2006.

[10] J. Schmidt, N. Hofemann, A. Haasch, J. Fritsch, and G. Sagerer, "Interacting with a mobile robot: Evaluating gestural object references," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2008, pp. 3804-3809.

[11] S. Sakurai, E. Sato, and T. Yamaguchi, "Recognizing pointing behavior using image processing for human-robot interaction," in *Advanced intelligent mechatronics, 2007 IEEE/ASME international conference on*, 2007, pp. 1-6.

[12] R. M. Holladay, A. D. Dragan, and S. S. Srinivasa, "Legible robot pointing," in *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, 2014, pp. 217-223.

[13] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin, "Generation and evaluation of communicative robot gesture," *International Journal of Social Robotics,* vol. 4, pp. 201-217, 2012.

[14] T. Spexard, S. Li, B. Wrede, J. Fritsch, G. Sagerer, O. Booij*, et al.*, "BIRON, where are you? Enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization," in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, 2006, pp. 934-940.

The final publication is available at Springer via http://dx.doi.org/10.1007/s12369-016-0348-9

[15] Y. Hato, S. Satake, T. Kanda, M. Imai, and N. Hagita, "Pointing to space: modeling of deictic interaction referring to regions," in *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, 2010, pp. 301-308.
[16] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press, 2004.
[17] I. Van Der Sluis and E. Krahmer, "Generating Referring Expressions in a Multimodal Context An empirically oriented approach," *Language and Computers,* vol. 37, pp. 158-176, 2001.
[18] S. L. Haywood, M. J. Pickering, and H. P. Branigan, "Do speakers avoid ambiguities during dialogue?," *Psychological Science,* vol. 16, pp. 362-366, 2005.
[19] I. Paraboni, K. van Deemter, and J. Masthoff, "Generating referring expressions: Making referents easy to identify," *Computational Linguistics,* vol. 33, pp. 229-254, 2007.
[20] A. Vaish and P. Kumari, "A Comparative Study on Machine Learning Algorithms in Emotion State Recognition Using ECG," in *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012*, 2014, pp. 1467-1476.
[21] P. Sharma and A. Vaish, "Information-Theoretic Measures on Intrinsic Mode Function for the Individual Identification Using EEG Sensors."
[22] A. Bangerter and E. Chevalley, "Pointing and describing in referential communication: When are pointing gestures used to communicate?," in *MOG 2007 Workshop on Multimodal Output Generation*, 2007, p. 17.
[23] A. G. Brooks and C. Breazeal, "Working with robots and objects: Revisiting deictic reference for achieving spatial common ground," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, 2006, pp. 297-304.
[24] A. C. Schultz and J. G. Trafton, "Towards collaboration with robots in shared space: spatial perspective and frames of reference," *Interactions,* vol. 12, pp. 22-24, 2005.
[25] D. F. Glas, T. Miyashita, H. Ishiguro, and N. Hagita, "Laser-Based Tracking of Human Position and Orientation Using Parametric Shape Modeling," in *Human-Robot Interaction in Social Robotics*, T. Kanda and H. Ishiguro, Eds., ed: CRC Press, 2012, pp. 158-236.
[26] J. K. Burgoon, D. B. Buller, J. L. Hale, and M. A. Turck, "Relational messages associated with nonverbal behaviors," *Human Communication Research,* vol. 10, pp. 351-378, 1984.
[27] D. L. Trout and H. M. Rosenfeld, "The effect of postural lean and body congruence on the judgment of psychotherapeutic rapport," *Journal of Nonverbal Behavior,* vol. 4, pp. 176-190, 1980.
[28] R. J. Edelmann, "The effect of embarrassed reactions upon others," *Australian Journal of Psychology,* vol. 34, pp. 359-367, 1982.
[29] J. M. Wolfe, "Guided search 2.0 A revised model of visual search," *Psychonomic bulletin & review,* vol. 1, pp. 202-238, 1994.
[30] S. Sternberg, "High-speed scanning in human memory," *Science,* vol. 153, pp. 652-654, 1966.
[31] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology,* vol. 12, pp. 97-136, 1980.
[32] A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth, "Deixis: How to determine demonstrated objects using a pointing cone," in *Gesture in human-computer interaction and simulation*, ed: Springer, 2006, pp. 300-311.
[33] P. Kühnlein and J. Stegmann, "Empirical issues in deictic gesture: referring to objects in simple identification tasks," *Report 2003/3, SFB,* vol. 360, 2003.
[34] J. M. Wolfe, T. S. Horowitz, N. Kenner, M. Hyle, and N. Vasan, "How fast can you change your mind? The speed of top-down guidance in visual search," *Vision research,* vol. 44, pp. 1411-1426, 2004.
[35] Q. Wang, P. Cavanagh, and M. Green, "Familiarity and pop-out in visual search," *Perception & Psychophysics,* vol. 56, pp. 495-500, 1994.
[36] D. F. Glas, T. Miyashita, H. Ishiguro, and N. Hagita, "Laser-Based Tracking of Human Position and Orientation Using Parametric Shape Modeling," *Advanced Robotics,* vol. 23, pp. 405-428, 2009.
[37] C. Müller, "Forms and uses of the Palm Up Open Hand: A case of a gesture family," *The semantics and pragmatics of everyday gestures,* pp. 234-256, 2004.
[38] S. Ishii, "Characteristics of Japanese nonverbal communicative behavior," *Communication (Journal of the Communication Association of the Pacific),* vol. 2, pp. 43-60, 1973.

[39] D. Richie, *A lateral view: Essays on culture and style in contemporary Japan*: Stone Bridge Press, 1998.

VITAE

Phoebe Liu received her S.B. degree in Electronic Engineering from Simon Fraser University, Canada in 2011, and received her M. Eng. in engineering science in 2013 from Osaka University, Osaka, Japan. She has been an internship researcher at the Intelligent Robotics and Communication Laboratories (IRC) at the Advanced Telecommunications Research Institute International (ATR) in Kyoto, Japan since 2011. She has been a Ph. D. candidate in the Graduate School of Engineering Science at Osaka University since 2013. Her research interests include machine learning for interaction design for human robot interaction and intelligent robots.

Dylan F. Glas received his Ph.D. in Robotics from Osaka University in 2013. He received his M.Eng in Aerospace Engineering from MIT in 2000 and S.B. degrees in Aerospace Engineering and in Earth, Atmospheric, and Planetary Sciences, also from MIT in 1997. From 1998-2000 he worked in the Tangible Media Group at the MIT Media Lab. He is currently a Senior Researcher and lead systems architect for the ERATO Ishiguro Symbiotic Human-Robot Interaction Project at Hiroshi Ishiguro Laboratories at the Advanced Telecommunications Research Institute International (ATR) in Kyoto, Japan. He is also a Guest Associate Professor at the Intelligent Robotics Laboratory at Osaka University. His research interests include social human-robot interaction, behavior design for autonomous robots, machine learning for social interaction, ubiquitous sensing, and teleoperation for social robots.

Takayuki Kanda received his B. Eng, M. Eng, and Ph. D. degrees in computer science from Kyoto University, Kyoto, Japan, in 1998, 2000, and 2003, respectively. From 2000 to 2003, he was an Intern Researcher at ATR Media Information Science Laboratories, and he is currently a Senior Researcher at ATR Intelligent Robotics and Communication Laboratories, Kyoto, Japan. His current research interests include intelligent robotics and human-robot interaction.

Hiroshi Ishiguro received a D.Eng. in systems engineering from the Osaka University, Japan in 1991. He is currently professor of Department of Systems Innovation in the Graduate School of Engineering Science at Osaka University (2009-) and distinguished professor of Osaka University (2013-). He is also group leader (2002-) of Hiroshi Ishiguro Laboratory at the Advanced Telecommunications Research Institute and the ATR fellow. He

was previously research associate (1992–1994) in the Graduate School of Engineering Science at Osaka University and associate professor (1998–2000) in the Department of Social Informatics at Kyoto University. He was also visiting scholar (1998–1999) at the University of California, San Diego, USA. He was associate professor (2000–2001) and professor (2001–2002) in the Department of Computer and Communication Sciences at Wakayama University. He then moved to Department of Adaptive Machine Systems in the Graduate School of Engineering at Osaka University as a professor (2002–2009). His research interests include distributed sensor systems, interactive robotics, and android science. Especially, his android studies are well-known very much in the world. He has received the Osaka Cultural Award in 2011.

Norihiro Hagita received the B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University in 1976, 1978, and 1986. In 1978, he joined Nippon Telegraph and Telephone Public Corporation (Now NTT). He was a visiting researcher in the Department of Psychology, University of California, Berkeley in 1989-90. He is currently Board Director of ATR and ATR Fellow, director of the Social Media Research Laboratory Group and the Intelligent Robotics and Communication Laboratories. He is the chairman of ATR Creative. He is also a visiting professor of Nara Institute of Science and Technology, Osaka University, and Kobe University. His major interests are cloud networked robotics, human-robot interaction, ambient intelligence, pattern recognition and learning, and data-mining technology. He has served as a chairman of technical committee in Network Robot Forum in Japan.