Improvement of object reference recognition through human robot alignment

Mitsuhiko Kimoto, Takamasa Iio, Masahiro Shiomi, Ivan Tanev, Katsunori Shimohara and Norihiro Hagita

***Abstract*—This paper reports an interactive approach to improve the recognition performance by robots of objects indicated by humans during human-robot interaction. We developed an approach based on two findings in conversations where a human refers to an object, which is confirmed by a robot. First, humans tend to use the same words or gestures as the robot in a phenomenon called alignment. Second, humans tend to decrease the amount of information in their references when the robot uses excess information in its confirmations: in other words, alignment inhibition. These findings lead to the following design; a robot should use enough information without being excessive to identify objects to improve recognition accuracy because humans will eventually use similar information to refer to those objects by alignment. If humans more frequently use the same information to identify objects, the robot can more easily recognize those being indicated by humans. To verify our design, we developed a robotic system to recognize the objects to which humans referred and conducted a control experiment that had 2 x 3 conditions; one factor was the robot's confirmation way and another was the arrangement of the objects. The first factor had two levels to identify objects: enough information and excess information. The second factor had three levels: congestion, two groups, and a sparse set. We measured the recognition accuracy of the objects humans referred to and the amount of information in their references. The success rate of the recognition and information amount was higher in the adequate information condition than in the excess condition in a particular situation. The results suggested the possibility that our proposed interactive approach improved recognition performance.**

I. Introduction

Social robots must provide services in real environments to recognize the objects indicted by humans as shown Fig. 1. To improve recognition performance, various approaches have been proposed. Nickel et al. used the 3D positions of a head and hands as well as the head's orientation to recognize pointing gestures in object references [1]. Schauerte et al. integrated speech and pointing gesture recognition by image processing [2]. Kemp et al. proposed a method that used a laser pointer to develop a new robotic interface so that people can easily indicate positions [3]. Since these works addressed the development of new devices or new algorithms, we describe them as engineering approaches, which is the most common type of works for improving recognition performance.

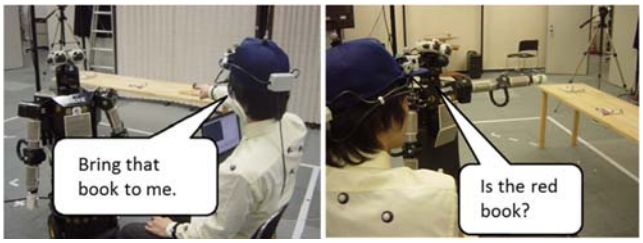

Figure 1. Robot recognizing an object indicated by a human

However, just engineering approaches will not solve the degradation of recognition performance in real conversations. For example, humans have enormous variability in their lexical choices in conversations [4]. Such variability degrades the recognition performance because humans might not always use the words contained in a database that stores object characteristics and they also do not always use enough words to identify an object [5]. Even if robots can perfectly recognize speech or pointing gestures, they might not distinguish an object indicated by humans from other objects.

In communication, humans solve such problems through the phenomenon of *alignment* with which humans tend to synchronize with their interlocutors such behaviors as vocabulary [6], syntax [7], lexical expressions [8], body movements [9,10] and facial expressions [11]. Through alignment, humans narrow down huge lexical choices and elicit terms, indications, or iconic gestures to naturally identify objects to their interlocutors.

The alignment findings in the interaction among humans inspired researchers to design behaviors of computers and robots to investigate alignment in human-computer or human-robot interactions to develop natural interfaces [12,13]. Since such works use interaction to improve computer or robotic systems, they are called interactive approaches. They look useful because they do not need special devices like engineering approaches. Moreover, they can be integrated in interaction designs independent of engineering approaches.

However, most works on alignment in human-computer or human-robot interaction have only investigated how alignment occurs without addressing how to apply it to real systems. Therefore, to the best of our knowledge, no research has reported how alignment should be applied to computer or robotic systems and whether it improves system performance.

*Research supported by the Strategic Information and Communications R & D Promotion Programme (SCOPE) and the Ministry of Internal Affairs and Communications (142107007)

M. Kimoto, I. Tanev, and K. Shimohara are with Doshisha University, Kyoto, 6100394, Japan (corresponding author to provide phone: +81-774-65-6949; e-mail: kimoto2013@sil.doshisha.ac.jp, {itanev, kshimoha}@mail.doshisha.ac.jp).

T. Iio, M. Shiomi, and N. Hagita are with ATR, Kyoto, 6190288, Japan. (e-mail: {iio, m-shiomi, hagita}@atr.jp).

In this paper, we developed a robotic system to recognize the objects to which humans refer and evaluated the effect of alignment in it.

II. Related Works

This section describes an overview of lexical and gestural alignment, which has been well studied by linguists, social and cognitive psychologists, and computer or robot scientists. Some works suggest that it is inhibited in certain situations.

A. Lexical alignment

In lexical alignment, two persons use the same terms for an object when they repeatedly talk about it [6-8]. Lexical alignment has been studied not only in human-human interaction but also in human-computer interaction [12,13] and human-robot interaction [14]. For example, Brennan suggested that humans readily adopted the terms of a computer partner through Wizard-of-Oz experiments using a database query task [12] and showed that the users of a spoken dialog system adapted their lexical choices to the system's vocabulary. Iio et al. conducted experiments in which a human referred to several objects in conversations with a robot. Their results revealed that humans tended to choose the identical terms and their categories used by the robot [14].

B. Gestural alignment

Gestural alignment has been observed where the speaker's gestures tend to synchronize to a partner's gestures in conversations. For instance, Charny reported that the postures of a patient and a therapist were congruent in psychological therapy [15]. Recent studies on embodied communication show that human gestures are entrained by robot gestures. Ogawa et al. developed a robot that synchronized its head nods with human speech. Through a conversation with a human, the entrainment of human nod motions was observed [16]. Ono et al. investigated human-robot communications involving giving/receiving route directions [17]. Iio et al. showed that people used more pointing gestures when a robot used gaze and pointing gestures [18]. Through entrainment, human gestures increased as robot gestures increased.

C. Alignment inhibition

Several studies reported cases where alignment became substandard in conversations. Shinozawa et al. investigated how humans referred to books when asking a robot to get them. Humans tended to use references with low information when a robot confirmed an indicated book using redundant information [5]. Holler and Wilkin found that mimicking co-speech gestures inhibited lexical alignment [23]. In their experiment, two interacting participants used both a verbal expression and a corresponding co-speech gesture at their first reference to an object, and then their word choice became less precise at their second reference despite consistent co-speech gestures. This phenomenon suggests that mimicking co-speech gestures is an integral part of establishing a shared understanding of referents and lexical alignment.

III. Interaction Design

A. Object reference conversation

Recognizing objects indicated by humans is one important ability for social robots. For example, when a human asks a robot to bring an object, she is referring to a specific thing. The robot must recognize it based on her speech or pointing gesture before getting it. Actually, robot systems for such interactions have already been extensively studied [20-22].

Therefore, we created an interaction called object reference conversation (Fig. 2). A robot asks a person to refer to an object in an environment where several objects are arranged (Ask). Next, she refers to an object (Refer), and the robot confirms the object to which she referred (Confirm). Then she answers whether the object confirmed by the robot is correct (Answer). Regardless of the answer, the conversation ends. In general cases, robots should continue to confirm objects until she also confirms, but we need to control the numbers of references by the human and confirmations by the robot. Our robot does not move to get the indicated objects because we focus on how humans refer to them.

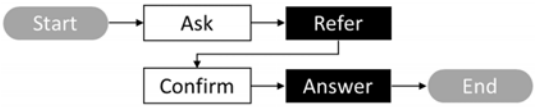

Figure 2. Object reference conversation: White and black boxes denote robot turns and human turns.

B. Confirmation behavior

We designed confirmation behavior to exploit alignment in object reference conversations because many related works observed alignment while a person repeatedly referred to an object and listened to confirmation interlocutors [6-8,12,14,17].

So that robots can precisely recognize objects indicated by humans, humans must use references that include enough information to identify the objects. The following summarizes the suggestions of related works about this requirement:

- Robots should utter necessary verbal expressions to identify an object because humans will come to use similar expressions. (Lexical alignment)
- Robots should point because humans will repeat that gesture. (Gestural alignment)
- Robots should avoid excessive verbal expressions because humans will decrease their own verbal expressions. (Alignment inhibition)
- Robot should avoid useless pointing gestures to identify an object because humans will decrease their own verbal expressions. (Alignment inhibition)

If humans provide enough verbal expressions to identify an object and point appropriately, it is easier for a robot to recognize the object indicated by humans. Since potential exists to improve the object recognition performance, we developed a robotic system for object reference conversations and designed confirmation behavior by a robot based on the above consideration.

IV. System

Fig. 3 illustrates the architecture of our developed system. When a human says something and points at an object, the speech recognition module extracts verbal expressions to identify the object from the speech, and the pointing gesture recognition module detects a pointing gesture and calculates its direction. The results of each module are associated in the integration module, which also calculates the likelihood of an object being referred to by the human among all objects. The system regards an object with the highest likelihood as the indicated object. Then the robot confirms the accuracy of the indicated object based on the recognition results. Confirmation is made in the confirmation selection module based on the recognition results, the object's information, and the other objects around it.

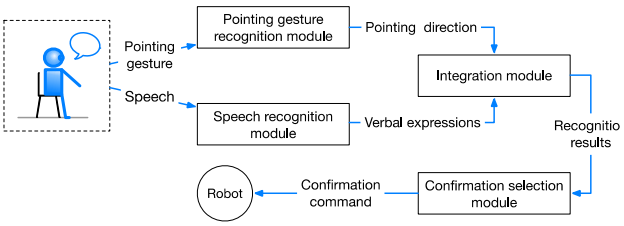

Figure 3. System architecture

A. Hardware

1) Robot

Robovie-R ver.2, which is a humanoid robot developed by the Intelligent Robotics and Communication Labs, ATR, has a human-like upper body designed for communication with humans. It has a head, two arms, a body, and a wheeled-type mobile base. On its head, it has two CCD cameras for eyes and a speaker for a mouth. The speaker can output recorded sound files installed on its internal controlled PC located in its body. We used XIMERA, which was developed in ATR, for speech synthesis. The following are the robot's degrees of freedom (DOFs): three for its neck and four for each arm. Its body has sufficient expressive ability to perform human-like gestures. It is 1100 mm high, 560 mm wide, 500 mm deep, and weighs about 57 kg. We developed greeting and asking reference motions with a motion development tool. These motions are played with synthesized speech when the robot first greets the human and asks her to specify the object.

2) Sensors

We use a small microphone and range image sensors for getting specification information. The microphone captures voices when humans specify an object and answer based on the robot's confirmation. It is attached to the human's body.

The range image sensor, which is called Kinect for Windows v2 and is developed by Microsoft, captures body frame data to recognize pointing gestures that refer to an object. The range image sensor is installed on the roof, 2.7 m from the floor. A web-camera is installed on the ceiling to detect objects in the environment through AR-markers. An external PC, which recognizes speech and pointing gestures, is also used as a database for the information of objects in the environment.

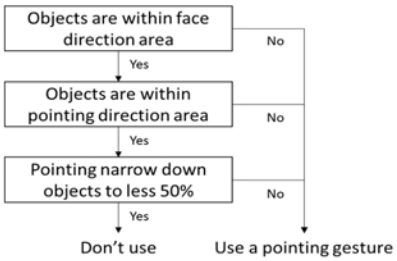

Figure 4. Determining whether to use a pointing gesture

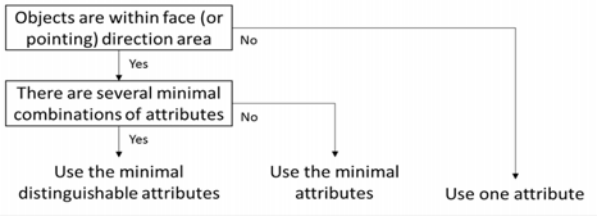

Figure 5. Selecting attributes of an object used by robot in confirmations

B. Software

The software contains four modules: speech recognition, pointing gesture recognition, integration, an object information database, and confirmation selection. First, voice and range images are sent to the speech and pointing gesture recognition modules, which calculate the reference likelihood of each object. Reference likelihood means the possibility that an object will be specified. The greater the reference likelihood, the higher is the possibility that humans indicated the object. The reference likelihoods calculated in the speech and pointing gesture recognition modules go to the integration module, which calculates the eventual reference likelihood and recognizes the indicated objects. The confirmation behavior selection module computes the minimum information for distinguishing the object from other objects using an object information database.

1) Speech recognition

The speech recognition module receives human speech that refers to an object and outputs each object's reference likelihood based on the speech recognition results. The likelihood is calculated as follows. First, the numbers of intersection factors among the attributes included in the speech and each object in the object information database are calculated and normalized. In our system, we used a speech recognition engine called Julius, which provides good performance in Japanese [19].

2) Pointing gesture recognition

The pointing gesture recognition module detects pointing gestures by using body frame data from the Kinect and calculates the reference likelihood of each object based on the pointing arm's vector. We modeled the likelihood as the difference from the pointing vector to a vector between a human and an object with a normal distribution function of $N(0, 1)$.

3) Integration

The integration module merges the reference likelihoods of the speech and pointing gesture recognition. The two likelihoods are summed and normalized. If participants don't use a speech or a pointing gesture, this module treats its

likelihood as zero. In summing and normalizing likelihoods of the speech and pointing gesture, we gives equal weight to the likelihoods in this system. This is because the accuracy of both speech and pointing recognition depends on situations, e.g. loudness of a speech, a speech rate, clarity of a pointing gesture and arrangement of objects, and so deciding a reasonable weight is difficult. The module sends the object id with the highest likelihood of other objects to the confirmation selection module.

4) Object information database

The object information database contains the attributes and the positions of all the objects in the environment. The attributes denote the verbal expressions used to identify objects, such as name, color, symbol, or shape. The attributes of objects are manually hand-coded. The object positions are automatically detected by AR-markers that are attached to the surface of every object [24].

5) Confirmation selection

The confirmation selection module receives the object id from the integration module and outputs a satisfactory confirmation without excess information to identify the indicated object. The module determines whether to use a pointing gesture and then selects the object's attributes uttered by the robot. We describe these processes in the following section.

a) With or without a pointing gesture

In this system, whether a pointing gesture is used depends on the extent to which the pointing gesture narrows down all of the objects to just that one confirmed by the robot. For example, if there are many objects, a pointing gesture does not narrow them down to the one confirmed by the robot; pointing gestures are not useful to identify one object out of many. Therefore, the robot does not use them in such cases.

The procedure of selecting whether to use a pointing gesture is shown in Fig. 4. First, the robot faces an object when confirming its selection. This face direction decides the area where the object exists. If there is only one object within that area, a pointing gesture can identify it. Even if there are other objects in the area, a pointing gesture can identify the object if it is alone within the area determined by the pointing gesture direction. In this case, the robot uses pointing gestures as well. If there are other objects in the pointing gesture's area, the decision whether to point depends on the extent to which the pointing gesture narrows down other objects from that object. Since pointing gestures, which narrow down other objects to less 50%, are useful to identify an object, we used them. In other cases, the robot did not use them.

b) Attributes used by robots in confirmations

The robot uses the minimal attributes of an object to identify it in a confirmation. Next we describe how to select the minimal attributes (Fig. 5).

First, the robot only gives one attribute that is chosen randomly if it confirms an object within an area decided by its face or pointing direction. In this case, only one attribute is sufficient to identify the object because a pointing gesture can distinguish it from the others. If there are other objects within the area, the robot uses enough minimal attributes to identify the object. If there are several sets of minimal attributes, we need to select one set from the other sets. In this case, the system calculates the similarity of the attributes in each set and chooses the set with the least similarity among the object and other objects, because we are considering a situation where speech recognition errors happen frequently. If the set of attributes is similar between the object and other objects, only missing one attribute causes failure of the object reference recognition. Therefore, not all of the sets of these objects should be similar.

To calculate the similarity of attributes, we used the Levenshtein distance of the letters of attributes. The Levenshtein distance is a string metric that measures the difference between two sequences. The greater the Levenshtein distance, the greater is the difference between two strings. The robot uses the minimal attributes with the highest Levenshtein distance among the object and other objects.

V. Evaluation

A. Trial design

We conducted an evaluation trial to investigate the effective of our developed system in several possible situations. We controlled the robot's confirmation behavior (confirmation factor) and the arrangement of objects in the environment (arrangement factor). In the following, we explain these controlled factors.

1) Confirmation factor

The confirmation factor had two levels: decent and excess. In the decent condition, the robot confirmed objects with enough information; the confirmations were based on our proposed approach. On the other hand, in the excess condition, the robot confirmed objects with excessive information. Therefore, the robot gave every attribute of an object and pointed during the confirmations.

The speech format of the confirmations is the following sequence of the attributes of objects:

[Deictic] [Figure] [Symbol] [Color] object name.

For example, the robot says, "That circle and red book?" or "That triangle, B and blue book?".

The confirmation factor was a within-participant condition.

2) Arrangement factor

We set this factor to check the influence of the arrangement of objects on the object reference recognition performance because the arrangement may affect what kinds

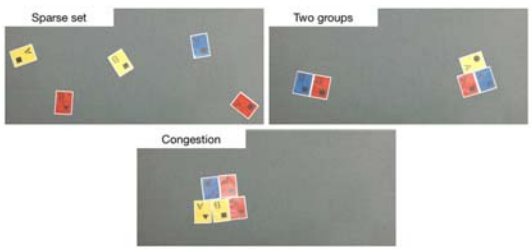

Figure 6. Examples of arrangement of books in each condition.

of verbal expressions and pointing gestures are chosen by people. The arrangement factor had three levels: congestion, two groups, and a sparse set. In the experiment, we asked participants to select books and arrange them freely under these three conditions. For example, we instructed them to "Stack the books close to each other," "Put books into three similar groups," and "Separate each book from the others" in the congestion, three groups, and sparse set conditions, respectively. Examples of a participant's arrangement are shown in Fig. 6.

B. Environment

The environment is shown in Fig. 7. We conducted our trial in a 1.5 m by 3.3 m rectangular area. The participants were seated in front of the robot. Five objects were placed between the robot and the participant. These objects are approximately 0.6-2.6 m far from the participants.

We controlled the attributes of the books. The size of the books was identical, and the attributes were color, figure, and symbol. Color, figure, and symbol contained three, three, and two types (Fig. 8). We prepared 18 books to satisfy all combinations of the attributes.

Figure 7. Environment

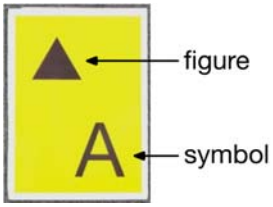

Figure 8. Example of object used in trial

C. Procedure

We conducted our trial as follows. First, we explained its purpose and our experiment's overview to participants who signed consent forms. After that we gave them the following oral instructions: "The robot can recognize human speech and pointing gestures. Please indicate one of the books after the robot asks you to do so. Please act toward the robot as you might toward a person." Besides participants were given these instructions with sitting on the chair placed in front of the robot.

After the instructions, the participant selected five books among the 18 and arranged them based on the arrangement conditions. The participants repeated the object reference conversations ten times. We call these procedures sessions, which were conducted in every arrangement condition: congestion, two groups, and sparse set. The participants answered questionnaires about their intention to use our system after three sessions. They eventually conducted two by three sessions with different confirmation conditions. We counterbalanced the order of the arrangement conditions within the sessions and the confirmation conditions within the trials.

D. Measurement

We measured the following items in the trials:

1) Recognition performance

Recognition performance denotes the success rate of the object reference recognition. We calculated it from the number of object references correctly recognized by the robot. We investigated whether the recognition performance is different between the decent and excess conditions.

2) Information content

Information content denotes the amount of information in the participant's references. The attributes of the objects (color, figure, and symbol) and the pointing gestures useful to identify them are regarded as information. We investigate whether the information content in the decent condition varies in the excess condition.

E. Participants

Eight native Japanese speakers participated in our experiment: six males and two females. They are all twenty-something university students.

VI. Results

A. Recognition performance

Table. 1 shows the recognition performance results, which we tested by a paired t-test to verify the effect of each factor. The recognition accuracy did not differ significantly between confirmation factors for any arrangement factor, though the recognition accuracy of the congestion level was higher in the decent condition than in excess condition about nine percent.

TABLE I. OBJECT RECOGNITION PERFORMANCE

	Sparse set	Two groups	Congestion	Total
Excess	73.8%	62.5%	61.3%	65.8%
Decent	72.5%	63.8%	70.0%	68.8%

B. Information amount

The results of the amount of information specification are shown in Table. 2. To verify the effect of each factor, we tested the results by a paired t-test. The information amount did not differ significantly between confirmation factors for any arrangement factor, nevertheless the information amount in the enough information level was higher than the redundant information level for all arrangement factors.

TABLE II. RESULT OF INFORMATION CONTENT[a]

	Sparse set	Two groups	Congestion	Total
Excess	1.50 (0.596)	1.59 (0.478)	1.91 (0.670)	1.67 (0.581)
Decent	1.56 (0.452)	1.78 (0.526)	1.94 (0.693)	1.76 (0.557)

a. Means with standard error in brackets

VII. Discussion and Conclusion

A. Interpretation of results

As a result of the experiment, there is no significant difference between the recognition accuracy and the amount of information for the confirmation factor. However they improved slightly in the proposed design. When the robot uses enough information instead of excessive information to identify objects, participants came to use similar information as the robot. But when it uses rich information for confirmation, participants used less information to identify objects. In this study, these tendencies were observed in differences in the performance of object reference recognition in the congestion level. We are set to verify these tendencies via further experiments. To verify the tendencies we need to raise the level of recognition accuracy because if the robot cannot recognize the participant's reference correctly, participants may feel the reference is not good and change the following references. In order to raise the recognition accuracy, it would be important to change the weight of speech and pointing gesture likelihood dynamically. In this study, we gave equal weight to the likelihoods independently of the arrangement of objects. Changing the weight in parallel with the objects arrangement (for example likelihood of a pointing gesture is given a larger weight when objects are put separately) would allow us to raise the recognition accuracy. Since improving the performance through interaction has never been reported in the field of human-robot interaction, we believe that a decent information approach will become a new way of designing robotic interaction.

Our design is that the performance of object reference recognition improves if robots make confirmations that contain minimum information for distinguishing objects. This is useful for designing interaction for social robots because confirmation is a natural behavior in object reference conversations and is easily integrated in interaction design. Since this approach is independent of such engineering restrictions as sensors and algorithms, it can be easily applied to existing robotic systems for object reference recognition.

B. Limitation

In this study, we experimented in a limited situation where participants specified books having three features: color, symbol, and figure. Our findings are general in the case of other objects, which also have such features as color, size, and shape that are useful for identification. The essence of our finding is that the information amount of robot confirmation must be controlled. Our findings will be observed in other object reference conversations.

References

[1] K. Nickel, and R Stiefelhagen, "Visual Recognition of Pointing Gestures for Human-Robot Interaction," *Image and Vision Computing*, vol. 25, no. 12, pp. 1875–1884, 2007.

[2] B. Schauerte, and G. A. Fink, "Focusing Computational Visual Attention in Multi-Modal Human-Robot Interaction," *Proc. 12th International Conference on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction*, pp. 1–8, 2010.

[3] C. C. Kemp, C. D. Anderson, H. Nguyen, A. J. Trevor, and Z. Xu, "A Point-and-Click Interface for the Real World: Laser Designation of Objects for Mobile Manipulation," in *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on*, 2008, pp. 241–248.

[4] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communications," *Communications of the ACM*, vol. 30, pp. 964–971, 1987

[5] K. Shinozawa, T. Miyashita, M. Kakio, and M. Hagita, "User Specification Method and Humanoid Confirmation Behavior," in *Humanoid Robots, 2007 7th IEEE-RAS International Conference on*, 2007, pp. 366–370.

[6] S.E. Brennan, and H. H. Clark, "Lexical choice and conceptual pacts in conversation, *Journal of Experimental Psychology: Learning*, Memory, and Cognition," vol. 22, pp. 1482–1493, 1996.

[7] H. P. Branigan, M. J. Pickering, and A. A. Cleland, "Syntactic coordination in dialogue," *Cognition*, vol. 75, B13–B25, 2000.

[8] S. Garrod, and A. Anderson, "Saying what you mean in dialog: A study in conceptual and semantic coordination," *Cognition*, vol. 27, pp. 181–218, 1987.

[9] A. E. Scheflen, "The significance of posture in communication systems," *Psychiatry*, vol. 27, pp. 316–331

[10] A. Kendon, "Movement coordination in social interaction: some examples described," *Acta Psychologica*, vol. 32, pp. 1–25, 1970.

[11] T. L. Chartrand, and J. A. Bargh, "The chameleon effect: The perception-behavior link and social interaction," *Journal of Personality and Social Psychology*, vol. 76, no. 6, pp. 893–910, 1999

[12] S. E. Brennan, "Conversation with and through computers," *User Modeling and User-Adapted Interaction*, vol. 1, no. 1, pp. 67–86, 1991.

[13] H. P. Branigan, M. J. Pickering, J. Pearson, and J. F. McLean, "Linguistic alignment between people and computers," *Journal of Pragmatics*, vol. 42, no. 9, pp. 2355–2368, 2010.

[14] T. Iio, M. Shiomi, K. Shinozawa, K. Shimohara, M. Miki, and N. Hagita, "Lexical entrainment in Human Robot Interaction," *International Journal of Social Robotics*, 2014, DOI: 10.1007/s12369-014-0255-x.

[15] E. J. Charny, "Psychosomatic manifestations of rapport in psychotherapy," *Psychosomatic Medicine*, vol. 28, no. 4, pp. 305–315, 1966.

[16] H. Ogawa, and T. Watanabe, "InterRobot: speech-driven embodiment interaction robot," *Advanced Robotics*, vol. 15, no. 3, pp. 371–377, 2001.

[17] T. Ono, M. Imai, and H. Ishiguro, "A Model of Embodied Communications with Gestures between Humans and Robots," in *Proc. of Twenty-third Annual Meeting of the Cognitive Science Society*, pp. 732–737, 2001.

[18] T. Iio, M. Shiomi, K. Shinozawa, T. Akimoto, K. Shimohara, and N. Hagita, "Investigating Entrainment of People's Pointing Gestures by Robot's Gestures Using a WOZ Method," *International Journal of Social Robotics*, vol. 3, no. 4, pp. 405–414, 2011.

[19] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, "Free software toolkit for Japanese large vocabulary continuous speech recognition," in *Proc. International Conf. on Spoken Language Processing (ICSLP)*, vol. 4, pp. 476–479, 2000.

[20] 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.

[21] D. O. Johnson, and A. Agah, "Human Robot Interaction Through Semantic Integration of Multiple Modalities, Dialog Management, and Contexts," *International Journal of Social Robotics*, vol. 1, no. 4, pp. 283–305, 2009.

[22] B. Schauerte, and G. A. Fink, "Focusing Computational Visual Attention in Multi-Modal Human-Robot Interaction," in *Proceedings of the 12th International Conference on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction*, pp. 8–12, 2010.

[23] J. Holler, and K. Wilkin, "Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue," *Journal of Nonverbal Behavior*, vol. 35, no. 2, pp. 133–153, 2011.

[24] H. Kato, and M. Billinghurst, "Marker Tracking and HMD Calibration for a Video-based Augmented Reality Conferencing System," in *Proceedings of the IWAR'99*, pp.85–94, 1999.