Robot Conversation Strategy for Indicated Object Recognition: Coordinating Alignment Mechanism and Gender Differences

by

Mitsuhiko Kimoto

Doctoral Dissertation

Submitted in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy in Engineering
in the Graduate School of Science and Engineering at
Doshisha University
Kyoto, Japan

March, 2019

Acknowledgments

First and foremost, I would like to thank my advisors at Doshisha University for their continuous support, guidance, and patience: Prof. Katsunori Shimohara and Prof. Ivan Tanev. Their depthful guidance always helped me deepen and broaden my research from various perspectives. Prof. Katsunori Shimohara patiently provided me with a lot of opportunities led me to grow as a research scientist. Prof. Ivan Tanev provided me with insightful suggestions that encouraged me to think deeply about research purposes, and his passionate attitude toward research inspired me to research harder. It has been a great honor to be their Ph.D. student.

I also would like to thank my dissertation co-advisor Prof. Masashi Okubo for his valuable advice on my research. Stimulating discussions with Prof. Masashi Okubo were invaluable in improving the dissertation.

My sincere thanks also goes to my advisors at Advanced Telecommunications Research Institute International (ATR): Dr. Masahiro Shiomi and Dr. Takamasa Iio. From my first day at ATR, they continuously provided me with excellent guidance about all aspects of abilities to be a good research scientist. Under their supervision, I actually got foundational experiences as a research scientist. Working at ATR has been an amazing experience, and Dr. Masahiro Shiomi and Dr. Takamasa Iio have been researcher role models for me.

Besides my advisors, I would like to thank the rest of my dissertation committee members: Prof. Miho Ohsaki and Prof. Seiji Tsuchiya for their insightful comments.

I am also grateful to Koya Kimura and Koki Ijuin. Thanks for personal and

academic interactions with them, I had fulfilling days during a Ph.D. student. In particular, I spent a lot of time studying with Koya Kimura from undergraduate days and had learned a lot from him.

My special gratitude also goes to the Japan Society for the Promotion of Science (JSPS) for helping and providing the funding for the research. I would like to also acknowledge anonymous paper reviewers for their critical comments that were essential to improve my research.

Last but not least, I would like to express my gratitude to my family. This dissertation would not have been possible without their understanding of my life as a Ph.D. student and selfless support. Thank You!

Mitsuhiko Kimoto
March, 2019

Abstract

This study proposes a system that employs a robot conversation strategy involving speech and gestures to improve a robot's indicated object recognition, i.e., the recognition of an object indicated by a human. We verify the usefulness and effectiveness of the proposed system from the perspectives of recognition performance and conversation impressions.

The progression of robotics has accelerated the research and development of social robots that provide services. For such robots to participate in human society, it is important that they have the ability to recognize objects indicated by humans. Indicated object recognition enables social robots to convey information about the objects and to pick up and transport the objects. Research conducted to improve the performance of indicated object recognition is divided into two main approaches: *engineering* and *interactive*. The engineering approach addresses the development of new devices or algorithms. Although such techniques improve the sensing capabilities of robots, recognizing objects indicated by humans remains difficult because human references to objects through speech alone are often ambiguous owing to the enormous lexical variability in human speech. Through human–robot interaction, the interactive approach improves the performance by decreasing the variability and ambiguity of the references.

Inspired by the findings of alignment and alignment inhibition, this study proposes a system that utilizes the interactive approach. While alignment is a phenomenon in which people use the same words or gestures as those of their

interlocutor, alignment inhibition is the opposite phenomenon in which people decrease the amount of information contained in their words and gestures when their interlocutor provides excess information. Based on these phenomena, we designed a robot conversation strategy in which a robot provides the minimum information needed to identify an object and uses pointing gestures to decrease the possible candidates of the referenced objects. In other words, the robot aligns its speech with that of humans, which contains useful information for identifying an object, and uses gestures considering alignment inhibition. As a result, the robot could elicit redundant references, and the performance of indicated object recognition could improve. Our system aims to incorporate as much valuable information as possible from humans to create alignments between robots and interlocutors to facilitate the identification of unique objects by the robots.

We thus developed a robotic system that uses combinations of speech, pointing gestures, and facial direction to recognize an object indicated by a human. Using a combination of recognition performance and conversation impressions, we experimentally compared this system with other interactive systems in which a robot explicitly requests clarifications when a human refers to an object. We also examined the gender differences of the alignment phenomena and analyzed the tendency of lexical alignment for a personal adaptive robot conversation strategy.

We obtained the following findings: (1) our system clarifies human references and improves indicated object recognition performance, (2) our proposed system forms better impressions than other interactive systems that explicitly request clarifications when people refer to objects, and (3) females align more with robots than do males.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Social Robots and Indicated Object Recognition

The progression of robotics technology has accelerated the development and research of social robots that interact with people and provide services. Social robots provide services, such as explaining exhibitions in a science museum [1], acting as shopping guides or assistants [2, 3], and caring for the elderly [4, 5, 6]. In addition, android robots that have a human-like appearance have recently been developed to engage people in natural communication [7, 8].

For social robots in human society, the ability to recognize an object referred to by users is important, as shown Figure 1.1. Such indicated object recognition enables social robots to provide information about the indicated object, convey the indicated object, and pick up the indicated object [9].

Research and development conducted to improve the performance of the indicated object recognition can be divided into two main approaches: *engineering* and *interactive*. The engineering approach addresses the development of new devices or new algorithms. Indicated object recognition consists of a wide range of component technologies, such as image processing, speech recognition, and natural language processing. Improvements to the component technologies will lead to a high accuracy recognition of an indicated object. 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

Figure 1.1 Robot recognizes an object to which a user referred.

references [10]. Kemp *et al*. proposed a method that uses a laser pointer to develop a new robotic interface enabling people to easily indicate positions [11]. Schauerte *et al*. used image processing to integrate speech and pointing gesture recognition [12], and Iwahashi *et al.* proposed a method of multi-modal language processing that learns the relationship between the users' lexical expressions and gestures and estimates the indicated object [13].

Even though such techniques improve the sensing capabilities of robots, recognizing the objects indicated by users remains difficult because user references are often ambiguous during conversations. People enjoy enormous variability in their lexical choices in conversations [14], which degrades the recognition performance because they might not always use the words contained in a database that stores an object's characteristics, and they do not always use enough words to identify an object [15]. Even if a robot can perfectly recognize speech or pointing gestures, they might not be able to distinguish an object indicated by humans from other objects.

The interactive approach improves the performance by decreasing such diversity and ambiguity of referencing through human–robot interaction. Based on the premise that humans refer to object in an ambiguous way, some researchers have proposed a method in which a robot explicitly asks users to provide additional

information to identify an indicated object [16, 17, 9]. For example, Hattori *et al.* proposed a system that integrates deep learning-based object detection and natural language processing to calculate the ambiguity of referencing and request additional information [9].

In contrast to these studies, Iio *et al.* proposed an approach that exploits alignments and clarifies the human's indication without explicit requests from the robots: the robot uses confirmation behavior to implicitly align with the people's indicating behaviors by eliciting lexical expressions contained in the robot's database from a user [18]. When people are talking, they tend to synchronize with an addressee such behaviors as lexical expressions [19, 20, 21], syntax [22], and body movements [23, 24, 25]. This phenomenon, known as alignment, occurs in interactions not only between humans but also between a human and artificial media such as spoken dialogue systems [26, 21, 27, 28, 29] and robots [30, 31]. Through alignment, humans narrow down huge lexical choices and elicit terms, indications, or iconic gestures to identify objects naturally for their interlocutors.

In this study, we take the stance of the interactive approach, which implicitly clarifies users' indications through human–robot interaction, and we bring up the absence of a perspective on the interaction between speech and gestures in the past research. While Iio *et al.* [18] considered how people use lexical information, they failed to consider how people use nonverbal information such as pointing gestures to recognize an indicated object. In other words, robots' gestures have not been considered when they interact with people. It is reported that robots' gestures affect humans' speech [15]; however, it remains unknown whether the robots' confirmation behavior considering both speech and gestures improves the performance of the indicated object recognition.

1.2 Research Objective

We propose a robot conversation strategy that elicit indications that are useful for identifying an indicated object from humans through the alignment of speech and gestures. Simultaneously, we evaluate the effectiveness of the proposed approach from the perspective of the recognition performance and impressions to humans. We also analyze the gender differences of the alignment for the design of a future personal-adaptive conversation strategy. This research answers the following three research questions:

1. Does a robot conversation strategy that exploits both lexical and gestural alignment improve the performance of indicated object recognition?

2. Which interactive robot conversation strategy, either robots explicitly request additional information or implicitly align with humans' indications, improves the performance of the indicated object recognition more and gives better impressions to humans?

3. Which gender is lexically entrained to robots, male or female?

1.3 Dissertation Organization

The remainder of this dissertation is organized as follows: In Chapter 2 summarizes the alignment phenomena and describes the design of the robot's conversation strategy to elicit clear referencing behavior from humans. Chapter 3 presents the system implementing the proposed strategy, describes the experiment conducted to verify the usefulness and effectiveness of the proposed strategy, and reports the results. In Chapter 4, we compare the proposed strategy with other interactive robot conversation

strategies, and we discuss the usefulness and effectiveness of the proposed strategy from the perspective of the performance and impressions of conversations. In Chapter 5 we analyze the gender differences of lexical alignment and report the tendencies for personal-adaptive conversation strategies. Chapter 6 presents the conclusions.

Chapter 2
Alignment Phenomena and Conversation Strategy

This chapter describes three types of alignment: lexical alignment, gestural alignment, and alignment inhibition, followed by robot confirmation behavior exploiting the three alignment types to improve the performance of indicated object recognition.

When communicating, humans tend to repeat lexical expressions that resemble those of their interlocutor [19, 21]. This phenomenon is called lexical alignment, and it is often associated with successful dialogues. Nenkova *et al.* [32] found that alignment in the use of high-frequency words correlated with task success and turn-taking in dialogues. Lee *et al.* [33] reported that the alignment measures of two prosodic features, pitch and energy, were higher in positive interactions between married couples than in negative interactions. According to Pickering and Garrod [34], alignment is a critical element for successful communication.

Alignment is also observed in interactions between a human and artificial media, for example, in spoken dialogue systems [28, 35, 26] and robots [31, 30, 36]. Iio *et al.* [30] found that lexical alignment and the alignment of word choices occur in conversations between humans and a robot. In their experiment, the participants were more likely to use the same words as the robot in conversations.

2.1 Lexical Alignment

In lexical alignment, two persons use the same terms for an object when they repeatedly talk about it [19, 22, 20]. Lexical alignment has been studied not only in human–human interaction but also in human–computer interaction [39, 35] and human–robot interaction [30]. For example, Brennan suggested that humans readily adopted the terms of a computer partner through Wizard-of-Oz experiments using a database query task [39] 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 [30].

These previous research studies indicate that humans tend to align their lexical expressions not only with their human interlocutors but also with artificial and/or robotic interlocutors.

2.2 Gestural Alignment

Gestural alignment has been observed where a speaker's gestures tend to synchronize with a partner's gestures in conversations. For instance, Charny reported that the postures of a patient and a therapist were congruent in psychological therapy [40]. Recent studies examining embodied communication show that human gestures are elicited by robot gestures. Ogawa *et al.* developed a robot that synchronized its head nods with human speech during a conversation with a human [41]. Ono *et al.* investigated human–robot communications involving giving and receiving route directions [42], and Iio *et al.* showed that people used more pointing gestures when

a robot used gaze and pointing gestures [31]. Through elicitation, human gestures increased as robot gestures increased.

The findings of these research studies indicate that, through alignment, human gestures increased as robot gestures increased.

2.3 Alignment Inhibition

Several studies reported cases where alignment became substandard in conversations. Iio *et al.* reported that humans tended to use references with low information when a robot confirmed an indicated book using redundant information [18]. Shinozawa *et al.* investigated how humans referred to books when asking a robot to get them. Their findings showed that humans tend to reference the object's attributes less than a robot when referring to an object [15]. Holler and Wilkin found that mimicking co-speech gestures inhibited lexical alignment [43]. In their experiment, two interacting participants used a verbal expression and a corresponding co-speech gesture in their first reference to an object; their word choice became less precise in their second reference despite consistent co-speech gestures. This finding suggests that mimicking co-speech gestures is an integral part of establishing a shared understanding of referents and lexical alignment.

These research studies indicate that humans tend to align their lexical expressions with their interlocutors less when the robot increases its use of lexical expressions in speech and gestures.

2.4 Robot Confirmation Exploiting Alignment Phenomena

To improve the performance of indicated object recognition, humans should use as much useful information as possible when referring to an object to uniquely identify the object based on it attributes, such as its color, form, and name. For example, if humans refer to an object with speech that contains many of the object's attributes, the robot's speech recognition could be robust to the failure of speech section detection and noisy speech. Additionally, if humans refer to an object using pointing gestures, the robot could narrow down the candidates of the indicated object based on the direction of the pointing. Hence, the desirable reference behavior to improve the performance of indicated object recognition is speech that contains as much useful information as possible to identify the object and pointing gestures (hereinafter known as a *redundant reference*). Considering the three abovementioned alignment phenomena, the following paragraphs summarize the approaches and their reasons to elicit a redundant reference from humans:

Lexical Alignment

- Humans tend to align their speech with their interlocutors. Therefore, robots should talk with useful lexical expressions to identify an object because humans will come to use the same or similar expressions.

Gestural Alignment

- Humans tend to align their gestures with their interlocutors and their gestures increase. Therefore, robots should use pointing gestures because humans will repeat the pointing gestures.

Alignment Inhibition

- If robots talk with many lexical expressions, humans tend to speak with fewer lexical expressions. Therefore, although robots should talk with useful lexical expressions to identify an object, they should avoid using too many lexical expressions because humans will decrease their use of lexical expressions in response.

- When aligning with robots' pointing gestures, humans tend to use fewer lexical expressions in their speech. Therefore, robots should avoid using pointing gestures in situations where the pointing gestures are not useful for identifying an object because humans will decrease their verbal expressions.

Therefore, to improve the performance of the indicated object recognition, robots should provide minimum information needed to identify an object and use pointing gestures to decrease the number of candidates for the object being referenced. In other words, robots should align their speech with that of humans, which contains useful information to identify an object, and use gestures considering alignment inhibition. Robots could thus elicit a redundant reference, and the performance of indicated object recognition could improve.

Chapter 3
Robot Confirmation Behavior Improving the Performance of Indicated Object Recognition

Chapter 3 verifies the effects of robot confirmation behavior exploiting alignment to improve the performance of indicated object recognition.

3.1 Interaction Design

To verify the effects of the robot's confirmation behavior exploiting alignment, this section describes the situation in which a human interacts with a robot.

Through an interaction called object reference conversations, Iio *et al.* verified the effectiveness of their robot's speech control, which exploited lexical alignment and its inhibition [18]. Such conversations focus on confirmation behavior, which is often observed in human–human communication. If a person cannot confidently understand which object is being referenced, they are likely to ask for confirmation. Furthermore, to avoid discrepancies in the interpretation, people sometimes confirm the referenced object even when it is clear.

This study examines the confirmation behavior in object reference conversations and verifies the effects of the robot's confirmation behavior exploiting the alignment phenomena.

Object reference conversations comprise four parts: *Ask*, *Refer*, *Confirm*, and

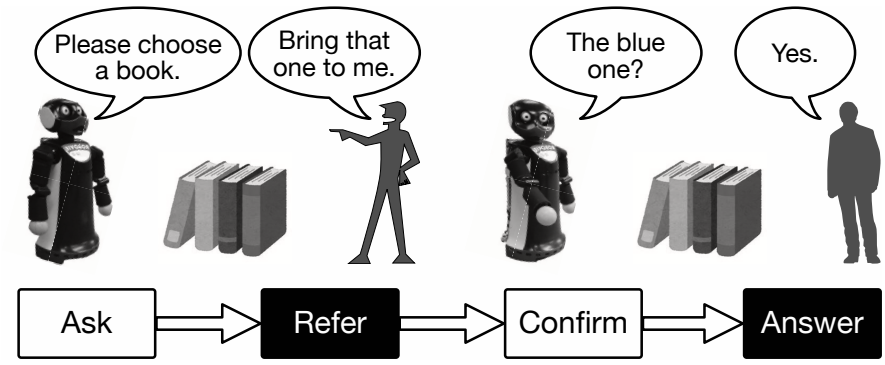

Figure 3.1 Object reference conversation: the white and black boxes denote the robot's and human's turns to speak, respectively.

Answer. First, a robot asks an interlocutor to refer to an object in an environment (*Ask*). Next the interlocutor refers to an object (*Refer*), and the robot confirms the object to which the interlocutor referred (*Confirm*). Then, the interlocutor answers whether the object confirmed by the robot is correct (*Answer*). Figure 3.1 shows the object reference conversation.

Section 3.2 describes how the system recognizes the object the human refers to in such an interaction between robots and humans.

3.2 System

Figure 3.2 shows the architecture of our developed system, which recognizes an object indicated by a user. The system comprises four parts: sensors, an indicated object recognition function, a confirmation behavior generation function, and an object information database. First, the system detects the positions and attributes of the objects arranged in the environment and saves them in the object information database before an object reference conversation. When a human refers to and points at an object, the speech recognition module extracts verbal expressions to identify the object from

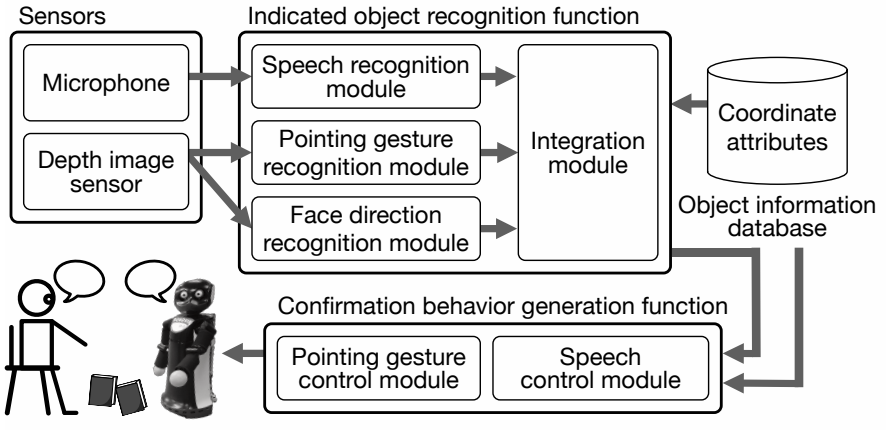

Figure 3.2 System architecture to recognize indicated object.

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 calculates the likelihood of an object being referred to by the human among all the objects. The system regards the object with the highest likelihood as the indicated object, and the robot confirms whether the estimated object is correct based on the recognition results. Confirmation is made in the confirmation behavior generation function based on the recognition results, the object's information, and other objects surrounding it. The confirmation behavior generation function is the proposed function exploiting the alignment phenomena to elicit a redundant reference from humans.

3.2.1 Robot

In this study, we used Robovie-R ver.2, a humanoid robot developed by the Intelligent Robotics and Communication Labs, ATR, which has a human-like upper body designed for communication with humans. The robot has three DOFs for its neck and four for each arm, and its body has an expressive ability for object reference conversations.

We used XIMERA for speech synthesis [44]. The robot is 1100 mm tall, 560 mm wide, 500 mm deep, and weighs about 57 kg.

3.2.2 Indicated Object Recognition Function

3.2.2.1 Speech Recognition Module

The speech recognition module receives human speech that refers to an object and outputs the normalized reference likelihood of each object based on speech recognition. To calculate the likelihood, the speech recognition module uses the number of attributes in the human speech, which is captured by a microphone attached to the human's collar. In this system, we used a speech recognition engine called Julius, which has a good performance in Japanese [45].

First, the speech recognition module extracts the attributes of a string from Julius and makes an attribute set of the speech R. Next, the module calculates the likelihood s of each object in the environment based on the number of shared attributes between the extracted attribute set and the attribute set in the object information database.

In the environment, depending on whether the number of n ($n \in \mathbb{N}^+$) objects are arranged, objects are arranged, the likelihood of the object h ($h \in \mathbb{N}^+$) based on speech recognition s_h is calculated as follows:

$$s_h = \frac{|O_h \cap R|}{\sum_{h=1}^{n} |O_h \cap R|} \tag{3.1}$$

where O_h indicates the attribute set of object h stored in the object information database.

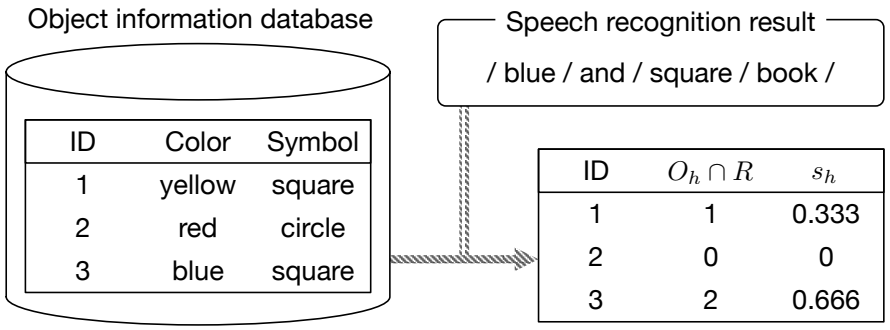

Figure 3.3 Example of the likelihood calculation based on speech recognition.

Figure 3.3 shows the example of the likelihood calculation based on the speech recognition. In this example, the speech recognition result is "/blue/and/square/book/" and the attribute set of this speech {blue, circle} is extracted. The number of shared attributes between the attribute set of speech and the attribute set of objects 1–3 in the object information database is 1, 0, and 2, respectively. Therefore, the likelihood based on speech recognition is calculated to be 0.333, 0 and 0.666, respectively relate to the objects 1–3.

3.2.2.2 Pointing Gesture Recognition Module

The pointing gesture recognition module calculates the normalized reference likelihood of each object based on pointing gesture recognition. The pointing gesture recognition module uses the body frame data from a depth image sensor called Kinect for Windows v2 and the object's position data stored in the object information database. We modeled the likelihood as the difference from the pointing vector, which connects a human head and the tip of the human hand, to a vector connecting the human head and an object with a normal distribution function $N(0, 1)$. The robot starts detecting pointing gestures when it asks an interlocutor sitting on a chair in front of it to refer to an object in the environment (*Ask*), and it finishes detecting when the interlocutor's reference speech is

recognized (*Refer*). When one hand or the other is more than 0.1 m vertically upward from the knee, the module recognizes the motion state as a pointing gesture. The module judges whether a pointing gesture is used according to the data obtained from the depth image sensor per 0.3 s. If a pointing gesture is detected, the module calculates the temporal likelihood of each object based on the data. After the detection, the module calculates the mean of the temporal likelihood for each object as the likelihood based on pointing gesture recognition.

On the k-th data with a pointing gesture, the temporal likelihood p_{hk} is defined as follows:

$$\alpha_{hk} = \arccos \frac{\boldsymbol{p}_k \cdot \boldsymbol{o}_{hk}}{|\boldsymbol{p}_k||\boldsymbol{o}_{hk}|} \tag{3.2}$$

$$g_{hk} = \frac{1}{\sqrt{2\pi}} \exp(-\frac{{\alpha_{hk}}^2}{2}) \tag{3.3}$$

$$p_{hk} = \frac{g_{hk}}{\sum_{h=1}^{n} g_{hk}} \tag{3.4}$$

Here, $\boldsymbol{p}_k$ and $\boldsymbol{o}_{hk}$ indicate the pointing vector and the vector connecting a human head and an object h, respectively, on the k-th data with a pointing gesture (Figure 3.4). g_{hk} indicates the probability that the object h is pointed at, and the angle α_{hk} between $\boldsymbol{p}_k$ and $\boldsymbol{o}_{hk}$ is defined as a random variable and modeled using the normal distribution function $N(0, 1)$ as g_{hk}.

In the detection section, when the temporal likelihood is calculated as m time, the likelihood p_h based on pointing gesture recognition is the mean of the temporal likelihood of each object during the section shown in Equation 3.5:

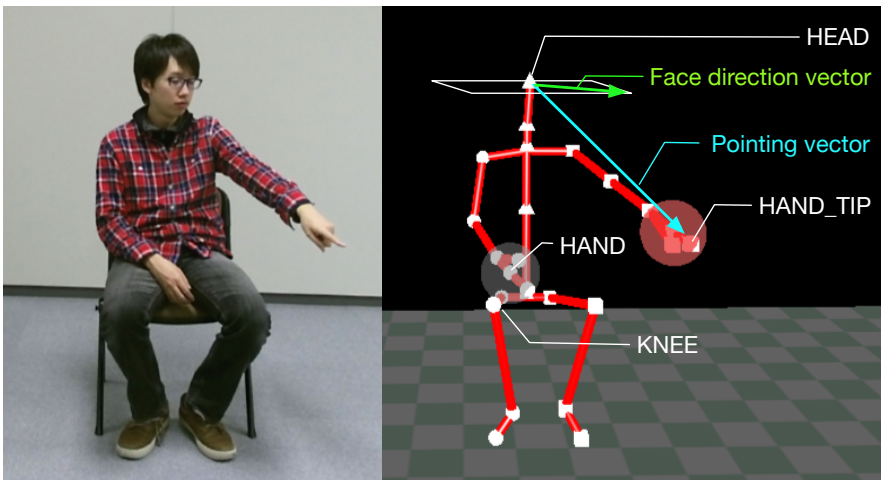

Figure 3.4 Joint information related to the pointing and face direction recognition.

$$p_h = \begin{cases} 0 & (m = 0) \\ \dfrac{\sum_{k=1}^{m} p_{hk}}{m} & (m > 0) \end{cases} \tag{3.5}$$

Figure 3.5 shows an example of the temporal likelihood calculation based on the pointing gesture recognition on the k-th data with a pointing gesture. First, the vector between a human head and object is calculated as $\boldsymbol{o}_{hk}$ according to the three-dimensional head position obtained from the depth image sensor and the positions of object 1–3 stored in the object information database. In addition, pointing vector $\boldsymbol{p}_k$ is calculated using the three-dimensional head position and the position of the tip of the pointing hand. The angle between $\boldsymbol{o}_{hk}$ of objects 1–3 and $\boldsymbol{p}_k$ is 0.340, 0.948 and 0.0639 rad respectively. As a result, the likelihood of the objects based on pointing gesture recognition on the k-th data is 0.366, 0.247, and 0.387, respectively.

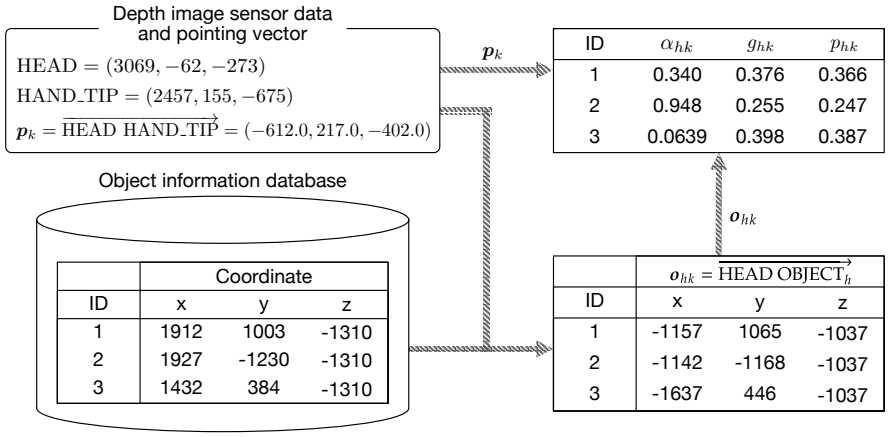

Figure 3.5 Example of the likelihood calculation based on the pointing gesture recognition.

3.2.2.3 Face Direction Recognition Module

The face direction recognition module calculates the normalized reference likelihood of each object based on face orientation data from the depth image sensor and the objects' position data stored in the object information database. The detection section for the face direction recognition is the same as that for the pointing gesture recognition. Thus, the module calculates temporal likelihood according to the data obtained from the depth image sensor per 0.3 s. The module then calculates the mean of the temporal likelihood of each object as the likelihood based on face direction recognition. The face direction recognition uses the angle β_f formed by the face orientation vector (Figure 3.4) on a level surface and the vector connecting a head and an object (Figure 3.6). The face orientation vector is calculated based on the head rotation angle θ ($0 \leq \theta \leq \pi/2$) rad obtained from the depth image sensor. If β_f is below $11\pi/18$ rad, the person is considered to be viewing the object; its likelihood is 1, and otherwise 0. This threshold is set because a humans' field of view is $11\pi/18$ rad at most [46, 47]. The likelihoods are normalized from 0 to 1.

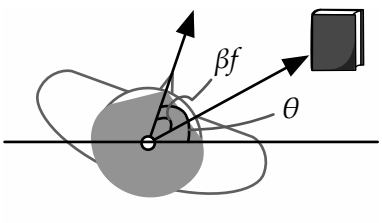

Figure 3.6 Angles used by the face direction recognition.

On the k-th data from the depth image sensor in the detection section, the object h's temporal likelihood f_{hk} is defined as follows:

$$f_{hk} = \frac{w_{hk}}{\sum_{h=1}^{n} w_{hk}} \tag{3.6}$$

$$w_{hk} = \begin{cases} 0 & (\beta_{hk} \geq \frac{11\pi}{18}) \\ 1 & (\beta_{hk} < \frac{11\pi}{18}) \end{cases} \tag{3.7}$$

where $\boldsymbol{f}_k$ indicates the face orientation vector on the k-th data. The angle β_{hk} between $\boldsymbol{f}_k$ and the vector connecting a head and an object is obtained in a similar way to Equation 3.2. Codomain of β_{hk} is $[0,\ \pi)$ because objects are in front of a person. The likelihood f_h based on face direction recognition is defined as follows:

$$f_h = \begin{cases} 0 & (j = 0) \\ \dfrac{\sum_{k=1}^{j} f_{hk}}{j} & (j > 0) \end{cases} \tag{3.8}$$

Figure 3.7 shows an example of the temporal likelihood calculation based on the face direction recognition. The two-dimensional vector $\boldsymbol{v}_{hk}$ connecting a head and an object is obtained based on a two-dimensional head position from the depth image

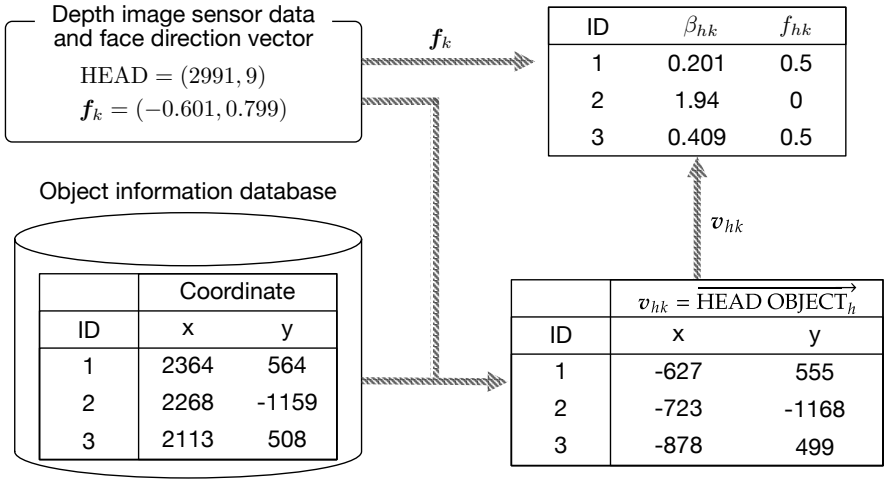

Figure 3.7 Example of the likelihood calculation based on the face direction recognition.

sensor and positions of objects 1–3 stored in the object information database. The face orientation vector f_k is obtained by a head rotation angle θ. In this example, β_f which is formed by f_k and v_{hk} of object 1–3 are 0.201, 1.94, and 0.409 rad, respectively. Only v_{2k} is not less than 110°, and the likelihood of each objects 1–3 are 0.5, 0 and 0.5, respectively.

3.2.2.4 Integration Module

The integration module merges the reference likelihoods of the speech, pointing gesture, and face direction recognition, and calculates the final likelihood. The final likelihood is obtained as the sum of the three likelihoods. In summing the likelihoods, we give equal weight to the likelihoods in this system because the accuracy of all speech, pointing, and face direction recognition depends on the situation, e.g., the loudness of a speech, a speech rate, the clarity of a pointing gesture, and the arrangement of objects; thus, deciding a reasonable weight is difficult.

The final likelihood of the object h is obtained as follows:

$$l_h = \frac{s_h + p_h + f_h}{\sum_{h=1}^{n}(s_h + p_h + f_h)} \tag{3.9}$$

The integration module obtains the object $o_{\max}$ with the highest likelihood of objects in the environment using Equation 3.10 and estimates the object as indicated object by a person.

$$o_{\max} = \arg\max_{x \in h}(l_x) \tag{3.10}$$

3.2.3 Confirmation Behavior Generation Function

The confirmation behavior generation function decides the robot's confirmation behavior by exploiting the alignment phenomena; that is, with or without a pointing gesture and the indicated object's attributes used for the robot's speech, following the approach described in Section 2.4.

3.2.3.1 Pointing Gesture Control Module

In this system, whether a pointing gesture is used depends on the extent to which how the pointing gesture narrows down the candidates for the indicated object. For example, if there are many objects adjacently, and a pointing gesture does not narrow the candidates for the indicated object, then pointing gestures are not useful for identifying one object out of many, and the robot does not use them in such cases.

The procedure of selecting whether to use a pointing gesture is as follows. First, we define the pointing and facing direction area centered on the indicated object, which is the area where the objects can be narrowed down by the robot's pointing gesture

and face direction, and the system calculates the number of objects existing in each range. As the definition of the pointing and facing direction area, we use the limit distance model proposed by Sugiyama *et al.* [48]. In the limit distance model, people cannot distinguish the indicated object if the edge of another object is in the area of θ_{L} from the indicated direction. In other words, the limit distance includes area θ_{L} and the distance from the center of another object to its edge. They reported that the limit angle of pointing gesture is $\pi/18$ rad through their experiment [48]. Accordingly, in this study, we decide the pointing direction area using the limit distance model and the limit angle. To determine the facing direction area, we used the limit distance using the limit angle as $\pi/9$ rad because the useful field of view, which is the visual area over which information can be extracted at a brief glance without eye or head movements [49, 50], is a maximum of $\pi/9$ rad [51]. If only one object is situated within the facing direction area, a pointing gesture can identify it. Thus, the robot confirms the indicated object with a pointing gesture. Even if other objects exist in the facing direction area, a pointing gesture can identify the object if it is alone within the pointing direction area. In this case, the robot confirms the indicated object with a pointing gesture as well. If there are other objects in the pointing gesture's area, the decision of whether to point depends on the ratio x between the number of objects in the facing direction area and in the pointing direction area:

$$x = \frac{\text{the Number of Objects in the Pointing Direction Area}}{\text{the Number of Objects in the Face Direction Area}} \tag{3.11}$$

In our study, the robot uses a pointing gesture in cases where $x < 0.5$. In other words, if the pointing gestures narrow down the candidates for the indicated object by 50%, the robot confirms the indicated object using a pointing gesture.

Figure 3.8 shows an example of how to decide whether to use a pointing gesture.

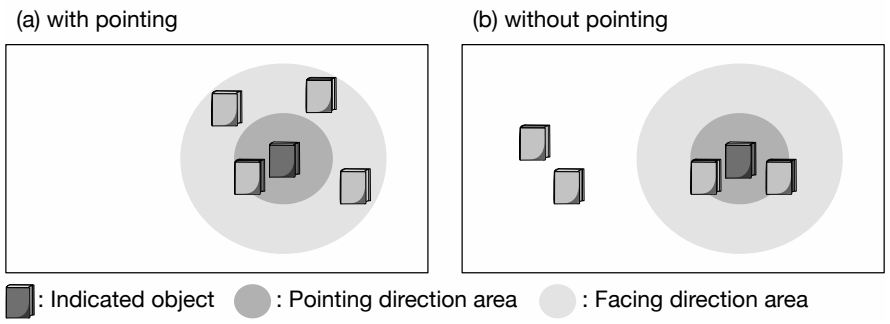

Figure 3.8 How to decide whether to use a pointing gesture.

In Figure 3.8a, there are two objects in the pointing direction area and five in the facing direction area, and the ratio $x = 0.4$. In such a case, pointing gestures are useful to narrow down the candidates for the indicated object, and the robot use a pointing gesture. In Figure 3.8b, there are three objects in the pointing direction area and three in the facing direction area, and the ratio $x = 1$. In such a case, pointing gestures do not narrow down the number of objects compared to the case in which the indicated object is pointed out only be the face direction, and the robot does not use a pointing gesture.

3.2.3.2 Speech Control Module

A robot uses the minimal number of object attributes to identify it in a confirmation. Next, we describe how to decide the attribute set used in the speech. First, the robot gives one attribute that is chosen randomly if it confirms an object within an area decided by the direction it is facing or pointing. In this case, one attribute is sufficient for identifying 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. Here, when confirming with a pointing gesture, the area means the pointing direction area, and when confirming without a pointing gesture, the area

is the facing direction area. If there are several sets of minimal attributes, we need to select one set. 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. If the object and other objects have similar sets of attributes, one missing attribute would cause a failure in the object reference recognition. similarity of the 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 the difference between two strings. The Levenshtein distance is defined as the minimum number of three edits—the insertions, deletions, or substitutions—required to transform one word into another. The robot uses the minimal attributes with the highest Levenshtein distance among the object and other objects. The Levenshtein distance between p-th character of stirng $\mathrm{s_{one}}$ and s-th character of stirng $\mathrm{s_{two}}$ is recursively given by $LD(p, s)$ as follows:

$$LD(p,s) = \min \begin{cases} LD(p-1,s)+1 \\ LD(p,s-1)+1 \\ LD(p-1,s-1)+c \end{cases} \tag{3.12}$$

$$c = \begin{cases} 0 & (char(\mathrm{s_{one}}, p) = char(\mathrm{s_{two}}, s)) \\ 2 & (char(\mathrm{s_{one}}, p) \neq char(\mathrm{s_{two}}, s)) \end{cases} \tag{3.13}$$

where c indicates the cost of substitutions, and $char(\mathrm{s}, i)$ is the function that indicates the i-th character of string s. We set the cost of the substitutions as two because substitutions can be expressed by deletions and insertions.

Figure 3.9 shows an example of how to decide enough minimal attributes to identify the object. In Figure 3.9, the minimum attribute sets to identify the indicated

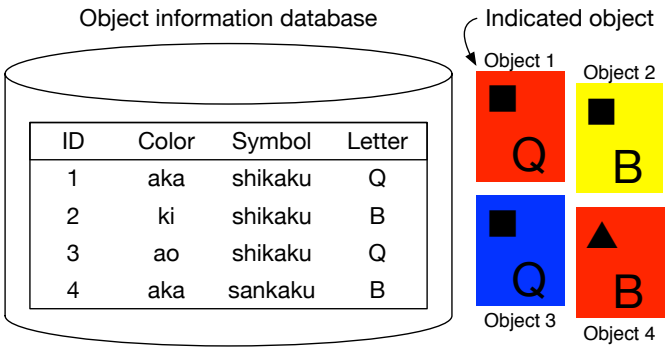

Figure 3.9 Selecting attributes of an indicated object for confirmation speech. The color aka, ki and ao mean red, yellow and blue respectively in Japanese. The symbol shikaku and sankaku mean square and triangle respectively in Japanese.

object are {aka, shikaku} or {aka, Q}. The minimum attribute set of objects 1 {red, square}, and the corresponding attribute sets of objects 2–4 are {ki, shikaku}, {ao, shikaku}, and {aka, sankaku}, respectively. In consequence, the Levenshtein distances of objects 2–4 are 3, 3, and 4, respectively, and the sum is 10. Similarly, the minimum attribute sets of objects 1 {aka, Q} and the corresponding attribute sets of objects 2–4 are are {ki, B}, {ao, Q}, and {aka, B}, respectively. Consequently, the Levenshtein distances of objects 2–4 are 8, 3, and 5, respectively, giving a sum of 16. Consequently, the attribute set {aka, Q} is more different from the attribute set of the other objects than the attribute set {aka, shikaku}, and the similarity of attributes is low. Therefore, in the situation shown in Figure 3.9 the attribute set {aka, Q} is used is used in the robot's confirmation speech.

3.3 Experiment

This section describes the experiment conducted to verify whether the robot confirmation behavior, exploiting both lexical and gestural alignment, improves the

performance of the indicated object recognition.

3.3.1 Hypothesis and Prediction

The desirable reference behavior of humans for robots to recognize an indicated object is the redundant reference, which refers to the human reference behavior and speech that contains as much useful information as possible to identify the object with pointing gestures. For the lexical and gestural alignment, if the robot confirms that an object with speech contains useful information for identifying the object accompanied by a pointing gesture, the human aligns with the robot's behavior and the robot can elicit the redundant reference from the humans. However, for the alignment inhibition, if the robot's speech contains much information, humans tend to decrease the amount of information in their speech. Similarly, if humans align with the robot's pointing gestures, they tend to decrease the amount of information contained in their accompanying speech. Therefore, the robot's confirmation with the speech containing much information that is valuable for identifying the indicated object with pointing gestures would be insufficient to elicit redundant reference behavior from humans. In other words, to elicit the redundant reference and improve the indicated object recognition, the robot needs to adjust the information contained in their confirmation behavior. Based on the discussions above and in Section 2.4, we make the following prediction:

Prediction: The performance of the indicated object recognition improves more in the case where a robot confirms an object using the minimum information needed to identify the object than when using all the useful information for identifying the object.

3.3.2 Conditions

To verify the hypothesis, we controlled the robot's confirmation behavior and the arrangement of objects in the environment. The factors and its levels are as follows:

Confirmation factor

- Minimum information level
- All information level

Arrangement factor

- Sparse set level
- Two groups level
- Congestion level

3.3.2.1 Confirmation factor

The confirmation factor was a within-participants design and had two levels: minimum information and all information. In the minimum information level, the robot confirms the indicated object with minimum information for distinguishing among several objects. In the all information level, the robot confirmed the objects using all the information available; thus, it gave every attribute of an object and pointed at each object during the confirmations. The speech format of the confirmations was the sequence of the object attributes. For example, the robot said, "That red book with a circle on it?" or "That blue book with a triangle and B on it?"

In this experimental design, for the confirmation factor, we did not separate the speech from the pointing gesture control to create a level because the content of the robot's speech is calculated depending on the result of the pointing gesture control, and

the strict separation of speech and pointing gesture control is difficult. In addition, even if the speech level is fixed to a minimum information level, it is difficult to distinguish the effects of a robot's gesture from the effects of a combination of speech and a robot's gesture. Therefore, in this experimental design, we did not separate the speech control from the pointing gesture control and treated both controls together as the minimum information level.

3.3.2.2 Arrangement factor

The arrangement factor was a within-participants design and had three levels: sparse set, two groups, and congestion. We set the arrangement factor to check the influence of the arrangement of the objects on the performance of the indicated object recognition because the arrangement might affect what kinds of verbal expressions and pointing gestures people choose. For example, if the objects were sparsely arranged, humans would refer to an object using pointing gestures and a deictic expression. However, if the objects were densely arranged, humans would refer to an object using a speech comprising many objects' attributes because pointing gestures alone are not enough to identify the object. Hence, not only the robot's confirmation behavior but also the arrangement of objects would influence the human's reference behavior, and the performance of the indicated object recognition would change depending on the arrangement.

In the experiment, we asked the participants to select books and arrange them freely under the following three conditions: "Arrange the books close to each other," "Arrange the books into two groups," and "Separate each book from the others" in the congestion, two groups, and sparse set conditions, respectively. Figure 3.10 shows examples of a participant's arrangement.

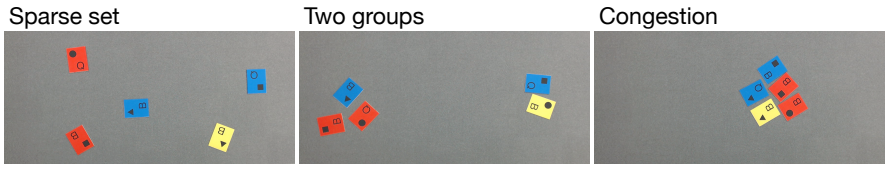

Figure 3.10 Examples of book arrangements in each condition.

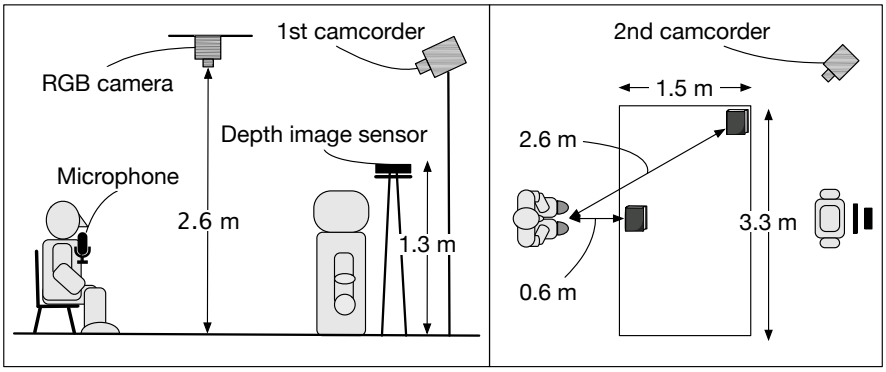

Figure 3.11 Experimental environment.

3.3.3 Environment

Figure 3.11 shows the experimental environment. The participants sat in front of the robot. The position of the robot and the chair were fixed to retain a distance between the robot and the participant and to eliminate the distance effects on the participant's object reference behavior. Objects were placed in a 1.5 m by 3.3 m rectangular area between the robot and the participant and grouped closely together without overlapping, approximately 0.6–2.6 m from the participants. We placed camcorders behind and on the right side of the robot to record the experiment (Figure 3.12).

As the objects the participants referred to, we used books that were 21 cm by 27.5 cm with attributes of color (red, blue, or yellow), a symbol (a circle, triangle, or square), and a letter on the cover (Q or B). We prepared 18 books to satisfy all the combinations of attributes. The attributes and positions of the arranged books were

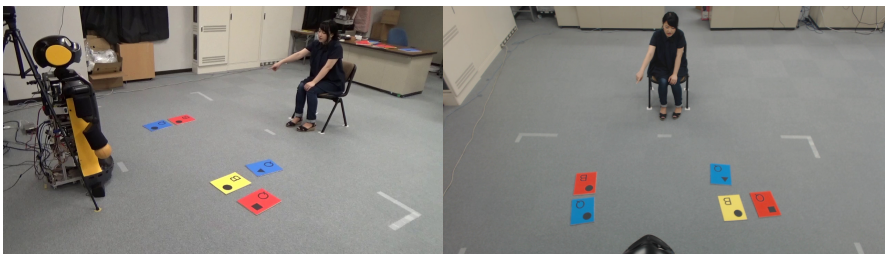

Figure 3.12 Example of an interaction scene in the experiment.

automatically recognized using an image obtained from a mounted RGB camera. The recognized attributes and positions were stored in the object information database.

3.3.4 Procedure

We conducted our experiment as follows. First, we explained the experiment to the participants and asked them to sign consent forms. Next, we gave them the following instructions orally: "The robot can recognize human speech, pointing gestures, and face directions. It will ask you to indicate a book. Do so as if you were speaking to another person." We did not give them instructions about how many times to refer to each object, the order of reference to objects, or what kinds of expression they could use for a reference.

After receiving the instructions, the participants selected five books among the 18 and arranged them according to the different arrangement levels. The participants repeated the object reference conversations 10 times to verify the influence of robot's confirmation behavior; it is necessary to hold the conversation multiple times because the robot's confirmation behavior affects a subsequent human's behavior. We decided the number of conversations based on related work that has verified human–robot alignment [18, 30]. An example of the conversation is as follows: First, the robot said, "Please choose a book," and the participants freely referred to an object (a book) in

the environment. The robot estimated the indicated object using its indicated object recognition function and confirmed the object by way of confirmation, calculated using the confirmation behavior generation function. After that, if the confirmed object was the same as the indicated object, the participant answered "Yes, it is" to the robot's confirmation. If the confirmed object was not the same as the indicated object, the participant answered "No, it isn't" to the robot's confirmation. The robot did not reply to the participant's answer.

We held 10 object reference conversation sessions, which were conducted in three arrangement levels: sparse set, two groups, and congestion. The participants eventually conducted six sessions, two confirmation levels by three arrangement levels. The participants selected five books and rearranged them at the start of each session, and thus there was spare time between conditions. We counterbalanced the order of the arrangement levels within the sessions and the confirmation levels within the trials.

3.3.5 Measurement

We measured the success rate of the indicated object recognition according to the success cases in which the book confirmed by the robot was the same as that indicated by a participant: the participant answered "Yes, it is" to the robot's confirmation. However, we also considered the cases in which the robot correctly confirmed the indicated object, but the participant answered "No, it isn't," and even though the robot mistakenly confirmed the non-indicated object, the participant answered "Yes, it is." The experimenter checked the videos from camcorders and the ceiling mounted RGB camera (Figure 3.12) and verified the recorded speech sound to identify such errors. Of 1,440 object reference conversations, there were only two error cases. In the first, although the robot confirmed the indicated object, the participant answered "No, it

isn't," and in the second, although the robot confirmed the non-indicated object, the participant answered "Yes, it is." We calculated these conversations as a success and error, respectively.

3.3.6 Participants

Twenty-four native Japanese speakers (12 females and 12 males, with an average age of 23.3 years, $SD = 7.61$) participated in our experiment.

3.4 Results

3.4.1 Verification of Prediction

Figure 3.13 shows the success rate of indicated object recognition. We conducted a two-factor repeated measure ANOVA and identified significant main effects in the confirmation factor ($F(1,23) = 4.916$, $p = .037$, $\eta_p{}^2 = .176$). This result showed that the success rate of the indicated object recognition with minimum information level was significantly higher than that at the all information level, thus supporting our prediction. We found no significant main effects in the arrangement factor ($F(2,46) = 2.245$, $p = .117$, $\eta_p{}^2 = .089$), and we found no significant interaction between the two factors ($F(2,46) = 2.659$, $p = .081$, $\eta_p{}^2 = .104$).

3.4.2 The Number of Object Attributes and Pointing Gestures in Humans' Indications

To investigate whether the number of object attributes and pointing gestures in the references changed depending on the robot's confirmation behavior, we measured the mean number of object attributes in the speech and pointing gestures in each

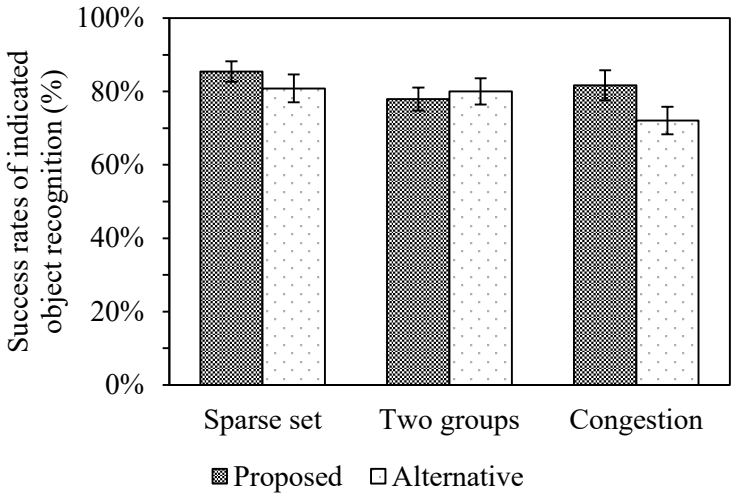

Figure 3.13 Performance of indicated object recognition with *SE*.

Table 3.1 Mean number of object attributes with *SE*.

	Sparse set	Two groups	Congestion
Proposed	1.5 (0.17)	1.6 (0.16)	1.7 (0.16)
Alternative	1.5 (0.17)	1.5 (0.16)	1.6 (0.16)

session. Table 3.1 and 3.2 show the mean number of object attributes and the mean number of pointing gestures, respectively. We defined the number of pointing gestures based on whether pointing gestures were used. If a participant referred to an object with pointing gestures, the number of pointing gestures was counted as one, and if a participant referred to an object without pointing gestures, the number of pointing gestures was counted as zero. We calculated the mean number of pointing gestures for each session, giving 10 object reference conversations. If a participant referred to an object with pointing gestures in 10 object reference conversations in one session, the mean number of pointing gestures was counted as one.

First, we conducted a two-factor repeated measure ANOVA for the mean

Table 3.2 Mean number of pointing gestures with *SE*.

	Sparse set	Two groups	Congestion
Proposed	0.69 (0.081)	0.68 (0.076)	0.65 (0.083)
Alternative	0.67 (0.079)	0.70 (0.081)	0.57 (0.094)

number of object attributes. As Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(2) = 16.8,\ p = .011$), the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.747$). We found significant main effects in the arrangement factor ($F(1.495, 34.384) = 5.026,\ p = .011,\ \eta_p{}^2 = .179$). Post hoc comparisons using the t-test with Bonferroni correction indicated a significant difference between the sparse set level and the congestion level ($p = .048$), and between the two groups level and the congestion level ($p = .035$). In other words, in the environment where objects were arranged close to each other, the participants tended to refer to an object with speech containing more attributes of an object than in the other environment of arranging books. The main effects in the confirmation factor were not revealed ($F(1, 23) = 1.120,\ p = .301,\ \eta_p{}^2 = .046$), and interactions between the confirmation and arrangement factor were also not revealed ($F(2, 46) = 1.000,\ p = .376,\ \eta_p{}^2 = .042$).

Next, we conducted a two-factor repeated measure ANOVA for the mean number of pointing gestures. The main effects were not revealed in the confirmation factor ($F(1, 23) = .956,\ p = .338,\ \eta_p{}^2 = .040$) or arrangement factor ($F(2, 46) = 2.775,\ p = .073,\ \eta_p{}^2 = .108$). In other words, a significant influence on the number of pointing gestures by the robot's confirmation behavior and the arrangement of objects was not observed.

3.4.3 Change of Referencing Style through Interaction

We proposed that the robot's confirmation behavior to elicit the redundant reference based on consideration of the human's desirable reference behavior to improve the performance of the indicated object recognition is the redundant reference, which is the reference behavior with the speech that contains as much useful information as possible for identifying the object and using pointing gestures, as described in Section 2.4. To verify whether the frequency of the such redundant reference depended on the robot's confirmation behavior, we analyzed the human's reference behavior from the viewpoint of the reference redundancy of speech, with or without pointing gestures.

First, the reference redundancy of speech is defined as the difference between the number of object attributes in the participant's references and the minimum number of attributes used for uniquely identifying the indicated objects in the environment. Our objects have three attributes (color, symbol, and letter), and the number of object attributes in the participant's references ranged from 0 to 3. For example, if a participant's reference has no attributes, the number of object attributes is 0. If a participant's reference contains all three attributes (i.e., a color, a symbol, and a letter), the number of object attributes is 3. The minimum number of attributes to uniquely identify the indicated object in the environment ranges from 1 to 3. Therefore, the reference redundancy ranges from -3 to 2. Therefore, if a participant refers to a book with no attributes (i.e., "that book" or "this book") in the environment where all the attributes are needed to uniquely identify the object (the minimum number of attributes is 3), the reference redundancy of speech is -3 because they failed to mention any of the attributes needed to identify the object. Likewise, if a participant refers to a book using all three attributes (i.e. a color, a symbol, and a letter) in an environment where there is only one red book and thus only the color attribute is needed to uniquely identify

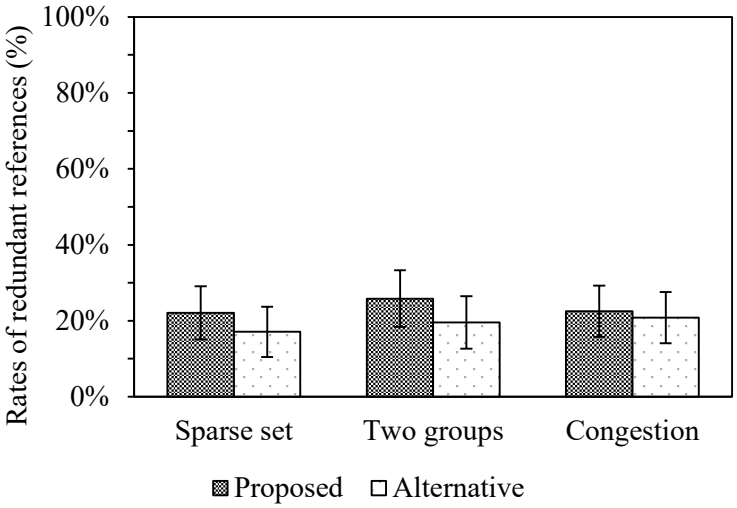

Figure 3.14 Rates of redundant references with *SE*.

the object (minimum number of attributes is 1), the reference redundancy of speech is 2 because they used two attributes more than necessary.

Next, we defined the redundant reference as the reference with the reference redundancy of speech more than 0 and with a pointing gesture, and measured the rate of the redundant reference per session, which is 10 object reference conversations. Figure 3.14 shows the results. We conducted a two-factor repeated measure ANOVA, and the main effects of the confirmation factor were close to significant ($F(1, 23) = 3.641$, $p = .069$, $\eta_p{}^2 = .137$). No effect was observed in the main effects of the arrangement factor ($F(2, 46) = .727$, $p = .489$, $\eta_p{}^2 = .031$) and interaction between confirmation $\times$ arrangement factor ($F(2, 46) = .312$, $p = .734$, $\eta_p{}^2 = .013$). These results suggest that if the robot confirms an indicated object with minimum information to identify the object by following the proposed method, humans tend to use the redundant reference.

3.5 Discussion

3.5.1 Implication

In this study, we verified whether the performance of the indicated object recognition improves through the robot's confirmation behavior. The experimental results show that the performance of indicated object recognition significantly improves when the robot's confirmation behavior contains the minimum information needed to distinguish the indicated object in the environment. The main contribution of this study is that we confirmed that the performance of the indicated object recognition could change only by changing the robot's confirmation behavior. Our contribution is useful for the design of human–robot interaction because confirmation behavior is often observed in human–human interaction, and the confirmation can be easily applied to the robot's behavior model to interact with people. In addition, the proposed method of changing the confirmation behavior does not depend on a speech recognition method, an algorithm for recognizing the reference behavior, or a variety of sensors, and could thus be easily applied to existing robotic systems for indicated object recognition.

3.5.2 Comparison with Confirmation Behavior Exploiting Only Lexical Alignment

This study does not make a direct comparison of the proposed confirmation behavior exploiting lexical and gestural alignment with the confirmation behavior exploiting only lexical alignment proposed by Iio *et al.* [18]. Therefore, we compared our proposed approach with past research from the viewpoint of the success rate of the indicated object recognition and the frequency of the redundant reference.

Table 3.3 shows the success rate of the indicated object recognition of Iio *et al.*'s

Table 3.3 Comparison of indicated object recognition performance with the past research work.

	Sparse set	Two groups	Congestion
Proposed	85%	78%	82%
Iio's approach	74%	78%	74%

work. First, we conducted a two-factor repeated measure ANOVA for the success rate of the indicated object recognition. First, we conducted a two-factor repeated measure ANOVA to determine the success rate of the indicated object recognition. The first factor is the exploiting approach factor that comprises two levels: the proposed level and Iio *et al.*'s approach level. The second level is the arrangement factor that comprises two levels, a sparse set and a congestion level. We skipped the two groups level because in Iio *et al.*'s experiment, the arrangement factor did not have two groups, but rather two or three groups. As a result of the ANOVA, no significant effects were observed for the exploiting approach factor ($F(1, 34) = 2.623,\ p = .115,\ {\eta_p}^2 = .072$), the arrangement factor ($F(1, 34) = .249,\ p = .621,\ {\eta_p}^2 = .007$), or the interaction between the exploiting approach × arrangement factor ($F(1, 34) = .249,\ p = .621,\ {\eta_p}^2 = .007$). Our proposed approach showed an improved success rate of the indicated object recognition. These results suggest that our proposed approach, which controls a speech and a pointing gesture immediately improves the success rate of the indicated object recognition and is more effective than Iio *et al.*'s approach, which only controls speech, although the direct comparison is difficult because the arrangement factor and algorithms of estimating an indicated object differ between the approaches.

Next, we analyzed the frequency of the redundant reference. In our study, it is suggested that if the robot confirms an indicated object using the proposed approach

of exploiting both a speech and pointing gesture, humans tend to confirm an object using a redundant reference. To analyze the frequency in Iio *et al.*'s experiment, we conducted a two-factor repeated measure ANOVA for the data of Iio *et al.*'s approach (the confirmation factor has two levels using a between-participants design and the arrangement factor has three levels using a within-participants design). Here, we remove the data of two participants, who arranged objects in three places, to coordinate the object arrangement with our experiment. However, no significant effects were observed for the confirmation factor ($F(1, 20) = .702$, $p = .412$, ${\eta_p}^2 = .034$), the arrangement factor ($F(2, 40) = 1.443$, $p = .248$, ${\eta_p}^2 = .067$), and the interaction between the confirmation $\times$ arrangement factor ($F(2, 40) = .1.477$, $p = .241$, ${\eta_p}^2 = .069$). In other words, in contrast to our proposed approach, Iio *et al.*'s work does not suggest that humans use a redundant reference to confirm an object.

These results suggest that the success rate of the indicated object recognition improves, and the change in reference behaviors that occurred through our proposed algorithm controlling a speech and a pointing gesture at a time compared with the past research work that only controlled a speech.

3.5.3 Influence of System Parameters on the Results

We decided the tolerance for the pointing gesture and the face direction based on related works. However, such tolerance would depend on the appearance of the robots and/or humans. For example, for a robot with a larger or smaller body size, we would need to change the line of sight and thus adjust the tolerance. Similarly, the parameters of Equation 3.11 might depend on the size of objects and the largeness of an environment. In each recognition module, we also decided the parameters, such as a field of view by referring related works. Such parameters have a certain level of

generality if a robot's conversation partner is human. However, such parameters might also need to be adjusted because of cultural differences, which would require changing the calculation method of the speech similarity in generating a confirmation speech; we used the Levenshtein distance in this study

3.5.4 Reducing Mental Burden During a Conversation

The conversation with the proposed approach would create a burden for people because the approach elicits references that include much information, even though the elicitation is implicit. To reduce the burden, the robot has a function that allows it to confirm an object in a way that implicitly allows people to use a simple reference if an indicated object recognition is evaluated as accurate, meaning there is no need to elicit references with more information based on the pre-calculated complexity of an environment and the accuracy of speech and gesture recognition. For example, in an environment with few candidates of an indicated object and a system that can always recognize an indicated object with high accuracy, the robot's confirmation with a deictic expression such as "That one?" or that uses object attributes used in a human's reference would not elicit much information and would thus reduce the mental burden during a conversation.

3.5.5 Limitations

We conducted our experiments in a limited situation, meaning that the participants referred to objects that have a limited number of attributes. In real environments, the attributes of objects are not limited and can influence the references. However, since the interaction manner between a robot and an interlocutor does not depend on the attributes of the objects, our findings can be generalized to other objects

In addition, we used objects that have a unique attribute combination and there were no same books. However, in a real environment, objects might have similar attributes, meaning that a an object cannot easily be distinguished from others, even if the speech contains all attributes of the object. In such a situation, to uniquely identify an object, the robots would need to provide an expression of a position relationship, i.e., "next to" or "near," to reduce the candidates of objects for a confirmation. Additionally, an expression giving a location in an environment, such as "in the corner of the room" or "on the table," would be useful to uniquely identify an object.

3.5.6 Conclusion

This chapter reports the use of an interactive approach in this study to improve the recognition performance of the objects a person refers to when speaking to a robot. We considered three phenomena in human–human and human–robot interaction to design the approach: lexical alignment, gestural alignment, and alignment inhibition. Alignment refers to a phenomenon in which people use the same words or gestures as those used by their interlocutor, and alignment inhibition is when people decrease the amount of information contained in their words and gestures when their interlocutor provides excess information. Based on these phenomena, we designed behavior policies, which stipulate that a robot should use sufficient information to identify objects, without being excessive, so that people would use similar information as the robot to refer to those objects, which would thus contribute to improved recognition performance. To verify our design, we developed a robotic system to recognize the object to which people referred, and we conducted an experiment in which we manipulated the amount of information used in the confirmation behavior. The results showed that the proposed approach improved the recognition performance of the

objects to which people referred.

Chapter 4

Conversation Strategy Comparison between Explicit Requests and Implicit Alignment in Object Reference Conversations

In this chapter, we experimentally compare two kinds of interaction strategies to decrease the ambiguity of references: implicit alignment (the proposed strategy described in Chapter 3) and explicit requests. In the first strategy, the robots implicitly align with the people's indicating behaviors. We call this the implicit alignment strategy. To encourage people to clarify their references, the robot can make an explicit request, by asking the person how to refer to the objects. For example, the robot explicitly asks a user, "Please describe the object's name, color, and size when pointing at it." The ambiguity of the user's references is expected to decrease if they fulfill the request.

However, it remains unknown which strategy more effectively decreases ambiguity and improves performance. Moreover, social robots should consider not only the recognition performance of the indicated objects but also the user's impression of the interaction. Even though the performance might be improved using a strategy, the performance gain becomes worthless if the users hesitate to interact with a robot by the chosen strategy and vice versa.

This chapter addresses whether the explicit request strategy or the implicit alignment strategy is better for object recognition contexts in conversations with

people. We developed a robot system that recognizes objects indicated by a user and experimentally compared the two strategies with our system. Based on the experiment results, we discuss the effectiveness of the two strategies.

4.1 Explicit Request and Implicit Alignment

The difference between the implicit alignment and the explicit request is whether a robot provides the specific information that is needed to recognize the objects and resolve the ambiguity of humans' reference behaviors. The following sections provide a detailed explanation of the implicit alignment and the explicit request.

4.1.1 Explicit Request

An explicit request is used to request the specific information the robot wants the interlocutors to use for recognizing the indicated object because it limits the references and reduces unexpected references. If an interlocutor refers to an object in line with the robot's instructions, the robot will likely recognize it with a high performance. A robot should also ask an interlocutor to make a reference that includes as much information as possible about the object that the robot can recognize. If the robot's recognition fails partly because of noise, insufficient speech volume, or unclear pointing gestures, references that include sufficient information increase the chances that the robot will correctly recognize the referenced object.

However, if the interlocutor does not follow the robot's requests, the robot should ask that they use all the requested information in the object references. This request reminds the interlocutor of the requests and encourages them to use all the information in subsequent conversations. Figure 4.1 shows an example of object reference strategy.

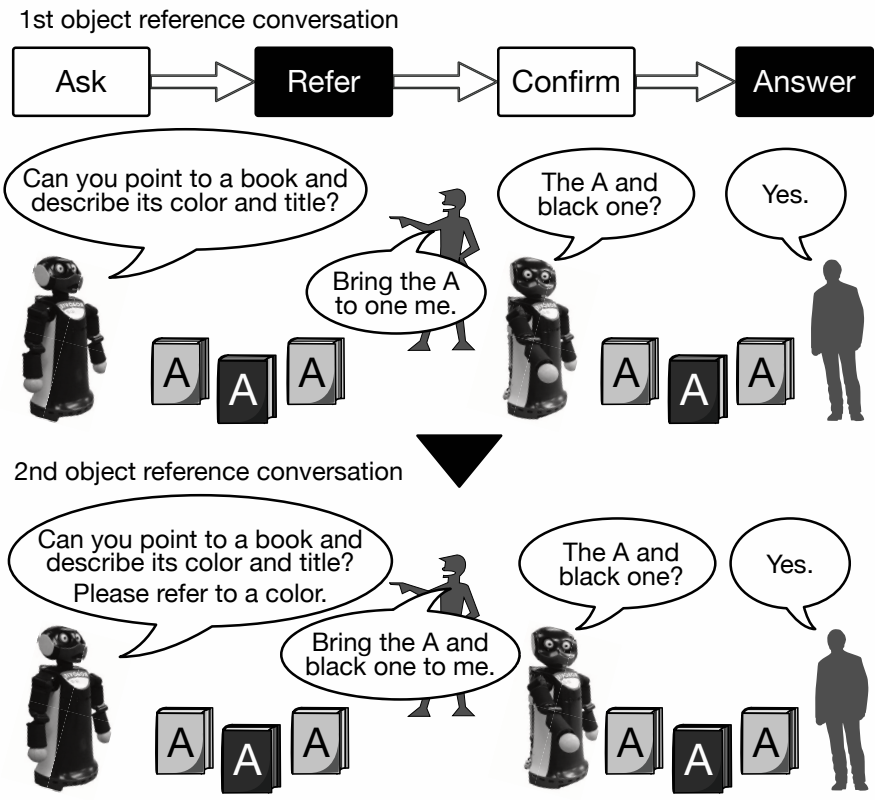

Figure 4.1 Example of the object reference conversation using the explicit request strategy.

Based on these considerations, we designed the explicit requests of the reference behavior as follows: A robot asks the interlocutors to make a reference that includes as much information as possible and requests that the interlocutor use any information that was missing from previous references.

4.1.2 Implicit Alignment

We adopted the implicit alignment strategy described in Section 2.4. This strategy exploits alignment in object reference conversations. Based on these three alignment phenomena—lexical alignment, gestural alignment, and alignment inhibition—we

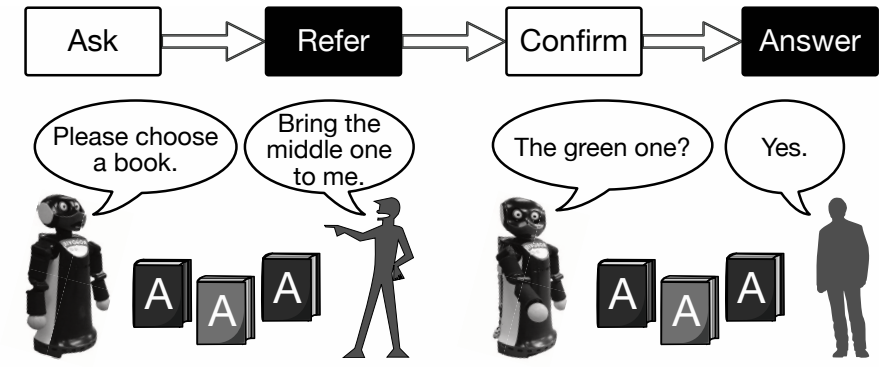

Figure 4.2 Example of an object reference conversation using the implicit alignment strategy.

designed robot behavior in which the robot uses the minimum information needed to distinguish among objects in the environment. Alignment inhibition is becoming substandard in conversations in some cases. However, through this design, humans learn to use references that include sufficient information to identify the objects by reducing the alignment inhibitions. This design was implemented in the *Confirm* part of the object reference conversation described in Section 3.1. In Chapter 3, our experimental results suggested the possibility of improving the recognition performance using the implicit alignment of reference behavior. However, despite discussing the effectiveness of reference behavior, we failed to evaluate the user's impression of the interaction. Therefore, we used this design of robot behavior as an implicit alignment of reference behavior; a robot should make confirmations that contain the minimum information needed for distinguishing among objects. Figure 4.2 shows an example of object reference conversations using an implicit alignment strategy.

4.2 System

Figure 4.3 shows the architecture of our developed system, which recognizes the objects indicated by a user. We developed the system by referring to past work that implemented the implicit alignment of reference behaviors for object reference conversations [37, 38]. The system comprises four parts: sensors, an indicated object recognition function, an object information database, and a conversation strategy selection function. Except for the conversation strategy selection function, the system is the same as that described in Section 3.2. First, the system detects the positions of the objects arranged in the environment and saves them in the object information database. When a user refers to an object, the indicated object recognition function recognizes the interlocutor's reference behavior and estimates the indicated object. The conversation strategy selection function chooses a robot behavior that corresponds to the implemented strategy and sends a behavior command to the robot. The robot confirms the indicated object and asks an interlocutor to refer to it in the next conversation in a way decided by the conversation strategy selection function.

4.2.1 Robot

In this study, we used Robovie-R ver.2, which is a humanoid robot developed by the Intelligent Robotics and Communication Labs, ATR, that has a human-like upper body designed for communication with humans. The speaker in its mouth can output recorded sound files from the internally-controlled PC located in its body. We used XIMERA for speech synthesis [44]. The robot has three DOFs for its neck and four DOFs for each arm, and its body has an expressive ability for object reference conversations. The robot is 1100 mm tall, 560 mm wide, 500 mm mm deep, and weighs about 57 kg.

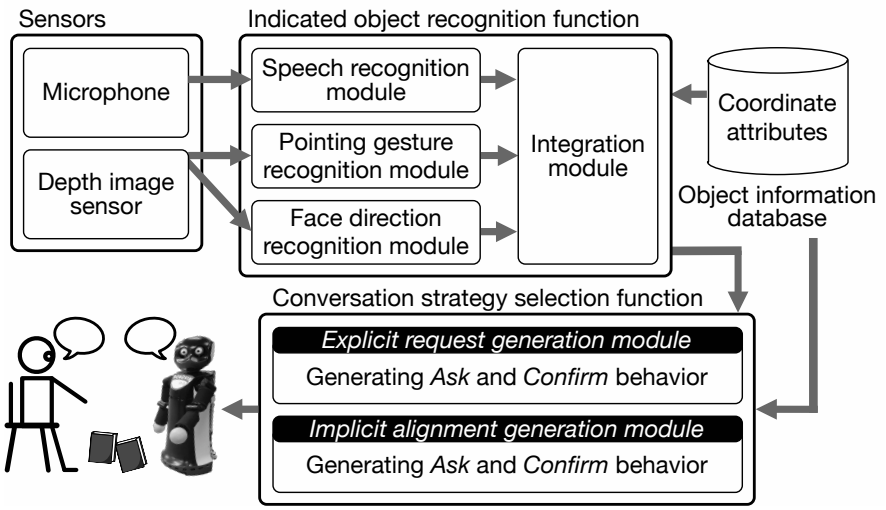

Figure 4.3 System architecture to recognize an indicated object.

4.2.2 Indicated Object Recognition Function

To develop this function, we implemented an algorithm [37, 38] that combines the results of speech recognition, pointing gesture recognition, and face direction recognition. The implementation of this function follows that described in Subsection 3.2.2, which we summarize in this subsection.

4.2.2.1 Speech Recognition

The speech recognition module receives human speech, which refers to an object and outputs the normalized reference likelihood of each object based on speech recognition. To calculate the likelihood, we used a previously proposed method [37, 38] that uses the number of attributes in the human speech, which is captured by a microphone attached to the human's collar. In this system, we used a speech recognition engine called Julius, which gives a good performance in Japanese [45].

4.2.2.2 Pointing Gesture Recognition

The pointing gesture recognition module obtains the body frame data from a depth image sensor called Kinect for Windows v2 and outputs the normalized reference likelihood of each object based on pointing gesture recognition. We modeled the likelihood as the difference from the pointing vector (between the human head and the tip of the human hand) to a vector between the human head and an object with a normal distribution function $N(0, 1)$.

4.2.2.3 Face Direction Recognition

The face direction recognition module obtains the face direction vector from the depth image sensor and outputs the reference likelihood based on the face direction recognition. We modeled the likelihood based on an angle parallel to the plane of the floor between the face direction vector and a vector between a human head and an object. If the vector is less than $11\pi/18$ rad, the person is considered to be viewing the object because a human's field of view is $11\pi/18$ rad at most [46, 47]; its likelihood is 1, otherwise 0. The likelihoods are normalized from 0 to 1.

4.2.2.4 Integration

The integration module merges the reference likelihoods of the speech and the pointing gesture and face direction recognitions. These three likelihoods are summed and normalized. The object with the highest likelihood is estimated to be the object indicated by the interlocutor.

4.2.3 Conversation Strategy Selection Function

The conversation strategy selection function determines how the robot confirms the indicated object (*Confirm* behavior) and how it asks an interlocutor to refer to an object (*Ask* behavior) in subsequent conversations. The conversation contents of the *Confirm* and *Ask* behaviors reflect whether the explicit request strategy or the implicit alignment strategy is used. The following sections describe in detail how to determine the robot's behaviors with each strategy.

4.2.3.1 Explicit Request Generation Module

When using the explicit request strategy, this module chooses the *Ask* behavior and adopts the explicit request strategy described in Subsection 4.1.1, and the robot explicitly provides requests about how to refer to the objects. The robot shows what kinds of information are needed for the robot recognition by providing an explanation to an interlocutor of the reference type, including all object attributes. A speech format of the explicit request has two parts. The first part is used every time, and the second part is only used when an interlocutor failed to follow the robot request and did not use all requested information in the last time reference. For example, the robot says, "Can you refer to a book by its color, the symbol on its cover, and the letter on its cover as well as by pointing and looking at it? Please refer to a color."

Figure 4.4 shows the procedure to generate the explicit requests. First, the explicit request generation module judges whether the object reference conversation is the first conversation. If it is the first the conversation, an explicit request is made to request a reference that includes object attributes that the robot can recognize, pointing gestures, and facing direction. If it is the second or a later conversation, the module judges whether the requested attributes and a pointing gesture were included in the

interlocutor's previous reference based on the speech and pointing gesture recognition results.

If an object o_p was indicated by an interlocutor in the previous conversation, and the reference likelihood based on a pointing gesture recognition is zero ($p_{o_p} = 0$), this means that an interlocutor referred to the object without a pointing gesture, and a sentence asking an interlocutor to use a pointing gesture in the following reference is added to the second part of the explicit request speech. Next, the module judges whether all requested attributes of the objects were included in the interlocutor's previous reference based on the speech recognition results. If an attribute is missing, the sentence asking an interlocutor to use the missing attribute in the following reference is added to the second part of the explicit request speech.

In the *Confirm* part, the robot confirms an object in the same way that it requested it from the interlocutor. Specifically, the robot confirms an indicated object o_{max} by saying all of the object's attributes $O_{o_{max}}$ and using a pointing gesture.

4.2.3.2 Implicit Alignment Generation Module

When using the implicit alignment strategy, the robot does not explicitly request how to refer to an object in the *Ask* part. The implicit alignment generation module generates a speech that asks an interlocutor to start referring to an object.

In the *Confirm* part, the robot in this module chooses the confirmation behavior and adopts the implicit alignment strategy, and the robot confirms the indicated object with minimum information for distinguishing among objects based on the design described in Subsection 4.1.2. By following the implementation described in Subsection 3.2.3, this module generates a confirmation behavior with the minimum information needed. The procedure comprises two steps: (1) deciding whether to use a pointing

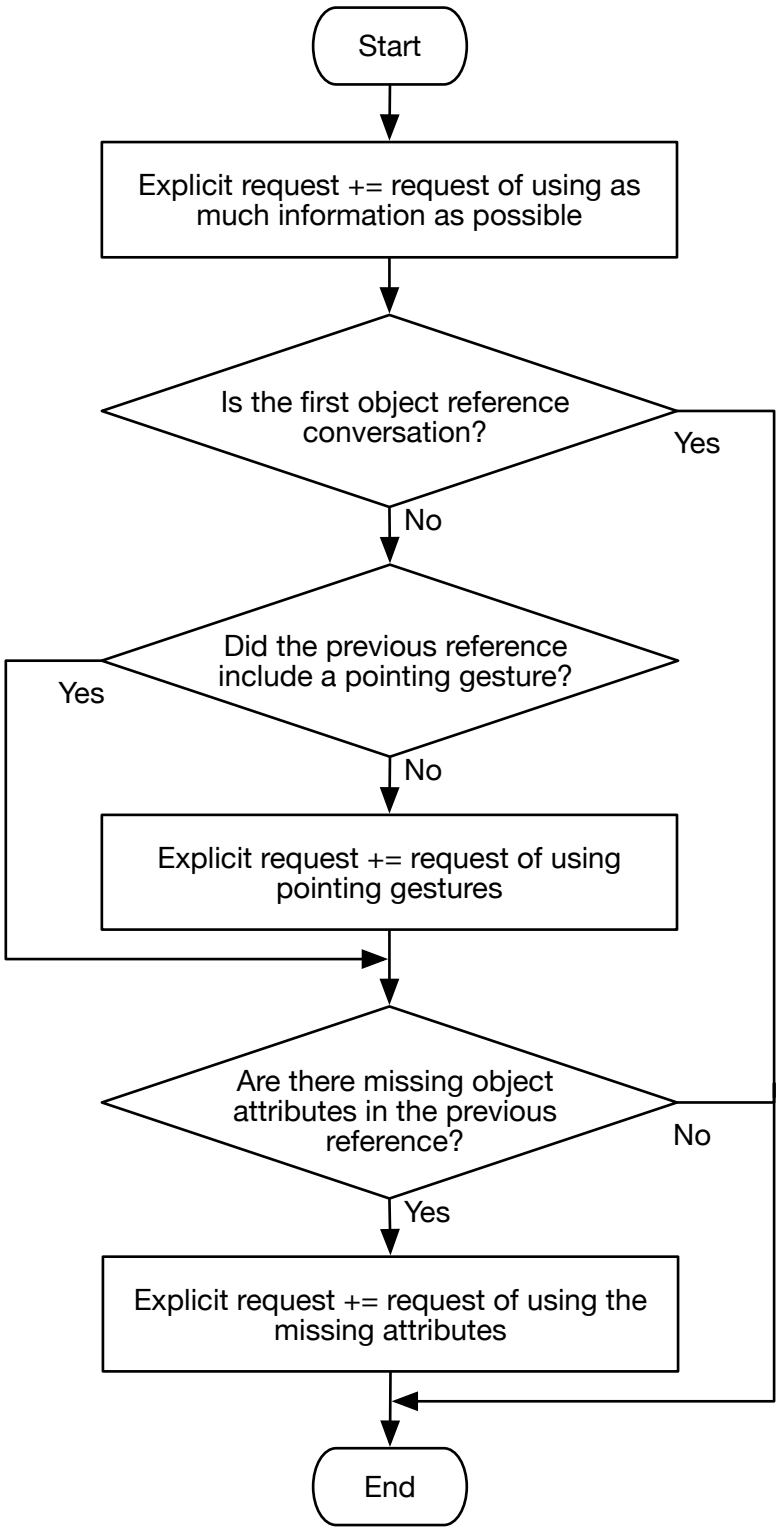

Figure 4.4 Procedure for making an explicit request.

gesture, and (2) deciding which object attributes to use in the confirmation speech. The details of this procedure are as follows:

First, this module decides whether to use a pointing gesture when the robot confirms an object; this choice depends on how the pointing gesture narrows down the candidates for the indicated object. For example, if there are many objects adjacently, a pointing gesture would not narrow the candidates for the indicated object; thus, pointing gestures are not useful for identifying one object out of many and the robot does not use them in such cases. In our study, if a pointing gesture narrows down the objects by 50%, the robot confirms the indicated object with a pointing gesture.

Second, this module decides (2) minimal attributes of an object to identify it from surrounding objects. If there is only one object within an area decided by its face or pointing direction area.the robot only gives one attribute that is chosen randomly. 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 set to identify the object.

For example, Figure 4.5a shows a situation using a pointing gesture: there are two objects in the pointing direction area, and the minimum attribute sets are “black and A” and “white and A”. The minimum attribute is “black”. The right side of Figure 4.5 shows a situation without using a pointing gesture: three objects are located in the facing direction area and the minimum attribute sets are “black and A,” “white and A,” and “black and Q”. The minimum attributes are “black and A”. The speech format of the confirmation is the sequence of attributes of the objects. For example, the robot says, “That white book with A on its cover?” or “That black book?”

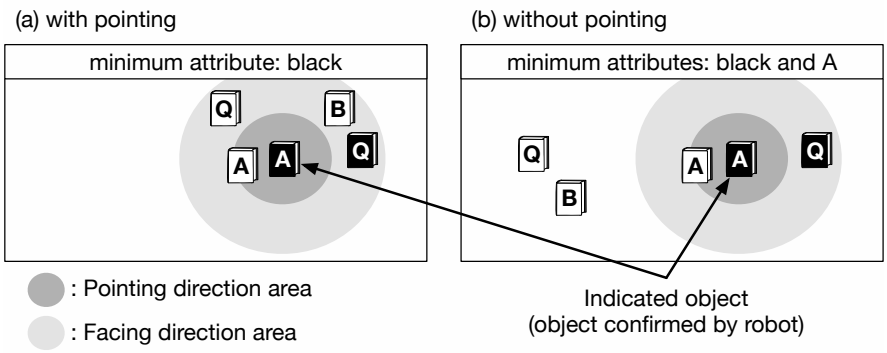

Figure 4.5 Minimum attributes of an indicated object.

4.3 Experiment

We experimentally compared the two interactive strategies: explicit request and implicit alignment.

4.3.1 Hypotheses and Predictions

If there is no explicit request about the reference type, the interlocutor might not know how to refer to an object, thus complicating the robot's ability to recognize indicated objects. While the interlocutor might feel the conversation is more natural than explicit requests, if the robot explicitly requests a particular reference type, the interlocutor knows how to refer to an object and may use it in object reference conversation. The references follow the robot's explicit request enable the robot to recognize the indicated objects more accurately. However, referencing an object based on explicit request is not common in daily conversations, and thus, the interlocutor might deem the conversation unnatural. Similarly, the interlocutor might feel the conversation creates a mental load and is troublesome because of the unaccustomed conversations. Based on these considerations, we make the following two predictions:

Prediction 1: The object reference recognition performance of conversations using the explicit request strategy will outperform that of conversations using the implicit alignment strategy.

Prediction 2: Conversations using the implicit alignment strategy will be perceived as having a lower mental load, and being less troublesome and more natural by the interlocutor than conversations using the explicit request strategy.

4.3.2 Environment

We used the same environment as shown in Figure 3.11, described in Subsection 3.3.3. The participants were seated in front of the robot. Five objects were placed in a 1.5 m by 3.3 m rectangular area between the robot and the participant. The books were grouped close together without overlapping. The objects were situated approximately 0.6–2.6 m from the participants.

We controlled the attributes of the books by following the past research work, which focused on object reference conversations [38]. All books were 21 cm by 27.5 cm. Their attributes were a color (red, blue, or yellow), a symbol (circle, triangle, or square), and a letter (Q or B) on the cover. We prepared 18 books to satisfy all combinations of attributes.

4.3.3 Conditions

We controlled a strategy that was applied to our developed system (applied strategy factor). The applied strategy factor had two levels: explicit request and implicit alignment. Both were respectively applied to the *Ask* and *Confirm* parts of the object reference conversation described in the interaction design section. The applied strategy

factor was a within-participant condition. In this experiment, we compared the method of the robot's speech and no difference was evident in the method used to recognize the interlocutor's reference behavior and estimate the indicated object.

In the explicit request condition, the robot asks interlocutors to make a reference that includes as much information as possible with comments that encourage the interlocutor to use the information missing in the *Ask* part. The speech format of the explicit requests includes two sentences. The first sentence is used every time, but the second sentence is only used when a participant does not use all of the information requested by the first sentence in the previous conversation. For example, the robot says, "Can you refer to a book using its color, a symbol on its cover, a letter on its cover as well as by pointing and looking at it? Please refer to a letter and point." In the *Confirm* part of this condition, since the robot confirmed the objects with all of the information, it gave every attribute of an object and pointed during the confirmations.

In the implicit alignment condition, unlike the explicit request condition, the robot does not explicitly provide requests about the reference type; it merely says, "Please choose a book" in the *Ask* part.

However, in the *Confirm* part, the robot utters a different sentence. For this purpose, we implemented an implicit alignment design for the reference behavior. In this condition, the robot confirms the object with minimum information to distinguish among the objects; the confirmations are based on the implicit alignment strategy of references proposed by Kimoto *et al.* [37, 38]. This strategy determines the robot's object reference behaviors, i.e., pointing behavior and speech, by considering the objects' position relationships and characteristics. The robot uses pointing gestures to decrease the number of candidates for the referenced objects. However, if the robot's pointing gesture becomes vague to an interlocutor, the robot does not use the pointing gesture.

The speech format of the confirmation is the sequence of attributes of the objects. For example, the robot says, “That blue book with a circle on its cover?” or “That yellow book?”

4.3.4 Measurement

4.3.4.1 Recognition Performance

The recognition performance is the success rate of the object reference recognition, which we calculated from the number of object references correctly recognized by the robot in each conversation session; each session was a set of 10 object reference recognitions.

4.3.4.2 Impression of Conversations

To investigate the participant’s impressions of the conversations, we prepared the following four questionnaire items and evaluated them using a seven-point scale ranging from 1 (disagree) to 7 (agree):

1. The conversation with the robot was a mental load (load feeling).
2. The conversation with the robot was troublesome (troublesome feeling).
3. The conversation with the robot was natural (natural feeling).
4. Overall impression of conversation (overall impression).

4.3.5 Procedure

We conducted our experiment as follows. First, we explained the experiment to the participants and asked them to sign consent forms. Next, we gave them the following

oral instructions: "The robot can recognize human speech, pointing gestures, and face direction. The robot will ask you to indicate a book. Please point it out as if you were addressing a person."

After the instructions, the participants followed the following steps:

1. Participant selects and arranges five books.
2. Participant has 10 object reference conversations under condition A.
3. Participant selects and arranges five books.
4. Participant has 10 object reference conversations under condition B.

Here, for the conditions A and B, we assigned the explicit request and implicit alignment, respectively. The assignment was counterbalanced. First, the participants selected five books from the 18 and arranged them according to the experimenter's instruction: "Please arrange the books in one place." We asked for the books to be arranged in one place because the recognition performance in the environment where the books were arranged was lower than that in the environment where books were arranged separately or in two places in the experiments that aimed to verify the effects of implicit alignment [38]. In addition, pointing gestures were used to a similar degree to those used in two groups and separate arrangements, which was about seven out of ten object reference conversations. We therefore considered that the one place arrangement is suitable for observing the change in the recognition performance through conversation strategies. Figure 4.6 shows an example of the arrangement. After placing the books in one place, the participants repeated the object reference conversations 10 times in both applied strategies (explicit request and implicit alignment). The participants answered questions about their impressions of the conversations after each conversation session.

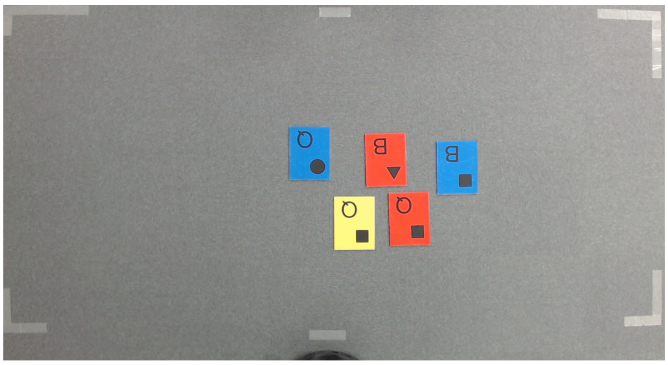

Figure 4.6 Example of book arrangements.

4.3.6 Participants

Twenty-six (13 females and 13 males) native Japanese speakers with an average age of 36.5 (SD = 9.3) participated in our experiment.

4.4 Results

4.4.1 Verification of Prediction 1

Figure 4.7 shows the recognition performance results. To verify the effect of each condition, we conducted a paired t-test and found no significant difference between the two interactive strategy conditions ($t(25) = -1.06$, $p = .302$, $d = .274$). This result indicated that Prediction 1 was not supported.

4.4.2 Verification of Prediction 2

Figure 4.8 shows the results of the questionnaire items. The load and troublesome feelings were reverse scored as an inverted scale, with seven treated as the most positive rating. To verify the effect of each condition, we conducted a paired t-test for each questionnaire item. Significant differences were found for the load feeling

Table 4.1 Examples of object reference conversations with explicit request and implicit alignment strategies.

		Ask	Refer	Confirm	Answer
Explicit	a)	Can you refer to the book by color, the symbol, and the letter on its cover as well as by pointing and looking at it? Please choose a book.	Well... fine, please take the blue triangle and B book.	That blue book with B and a triangle?	Yes, it is.
	b)	Can you refer to the book by color, the symbol, and the letter on its cover as well as by pointing and looking at it? Please refer to a color. Please choose a book.	Yellow book with a triangle.	That yellow book with a triangle and B?	Yes.
Implicit	a)	Please choose a book.	Hmm, please take the red book with B.	That red book with B?	Yes, it is.
	b)	Please choose a book.	Well, please choose that red book with Q and a circle.	That book with a circle?	Yes, it is.

($t(25) = 2.440$, $p = .022$, $d = .58$), for the troublesome feeling ($t(25) = 4.556$, $p < .001$, $d = .89$), for the natural feeling ($t(25) = -2.403$, $p = .024$, $d = .51$), and for the overall impression among the conditions ($t(25) = -2.339$, $p = .028$, $d = .36$). These results supported Prediction 2.

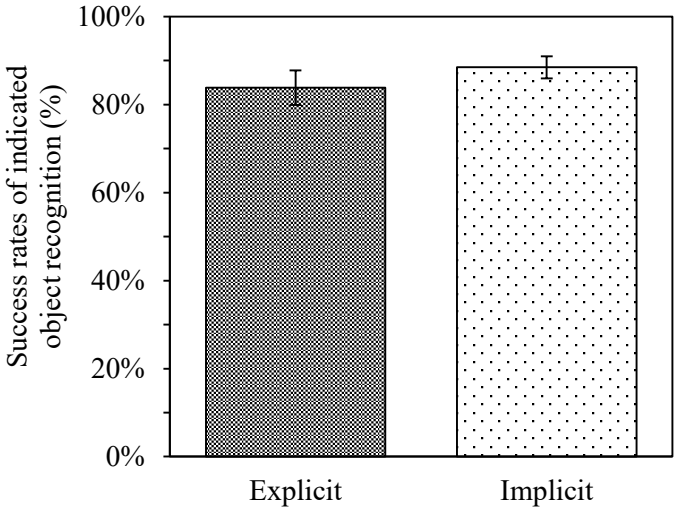

Figure 4.7 Performance of indicated object recognition with *SE*.

4.5 Discussion

4.5.1 Comparison between Explicit Request and Implicit Alignment

No significant difference was identified in the recognition performances between the two strategies. This result can be interpreted in two ways. First, we consider that the implicit alignment strategy improved the recognition performance at the same level as the explicit request strategy. Although we cannot verify the effects of the implicit alignment strategy in our experimental settings, a counter condition of an implicit alignment strategy is not a no-implicit alignment strategy but an explicit request strategy, and the implicit alignment strategy likely improves the recognition performance, as argued by past work [38, 18].

Second, the explicit request strategy might not improve the recognition performance very much. We predicted that the interlocutors would refer to an object as the robot instructed, but in the experiment, 28% of references in the explicit request

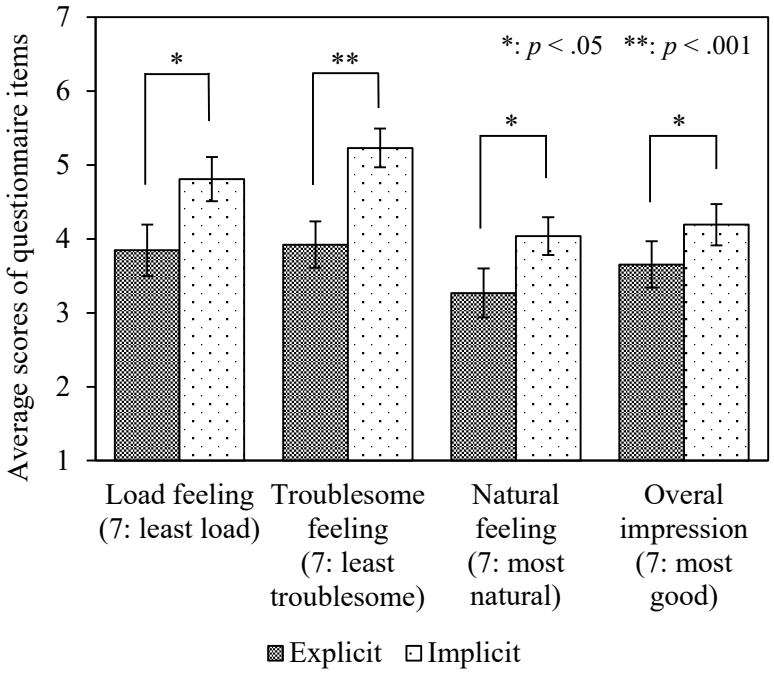

Figure 4.8 Impressions of conversations with *SE*.

condition did not follow the robot's requests. Some instructed with information that was dropped such as a pointing gesture and a color. The explicit request is likely ineffective for inducing interlocutors to encourage users to adopt clear but recognizable references.

The experiment results show that the interlocutor's impressions of the conversations with the implicit alignment strategy were perceived to have a significantly lower mental load and to be less troublesome and more natural than the explicit request strategy. The overall impression of the conversations with the implicit alignment strategy is also rated as higher than that of the conversations with the explicit request strategy. Accordingly, the explicit request tends to be unnatural for people and creates feelings of uneasy interaction among the interlocutors. In the conversations containing an explicit request, if an interlocutor does not follow the

Table 4.2 Correlation between the number of additional requests and conversation impressions.

	Load feeling	Troublesome feeling	Natural feeling	Overall impression
Pearson's r	−0.040	0.034	0.169	0.072
p-value	0.847	0.870	0.408	0.727

robot's request and refer to an object with missing information, the robot requests the missing information in the next *Ask* part. Such additional requests might affect the impression of the conversations. Therefore, to analyze the correlation between the numbers of additional requests and the impressions of the conversations, we conducted a correlation analysis using Pearson's correlation. Table 4.2 shows the results. We found no correlation between the two variables, which suggests that the explicit request from the robot to humans is perceived to be a mental load, troublesome, and unnatural by itself.

We therefore conclude that the implicit alignment strategy is better than the explicit request strategy for object recognition contexts in conversations with people. Our findings are useful for designing interactions for social robots; good impressions of conversations are important for them because they interact often with people. These findings can be integrated in object recognition contexts and many other contexts since determining whether to use an explicit or implicit strategy is conceivable in other contexts.

4.5.2 Relationship between Personality and Impression of Conversation

The personality of humans might affect the impressions of conversations. To examine the effects of the personality on the impressions of the conversations, we measured the participants' personalities using a personality scale of the Big Five personality traits,

known as the five-factor model. The five factors are labeled as follows [52]:

- Extraversion (talkative, assertive, energetic)
- Agreeableness (good-natured, cooperative, trustful)
- Conscientiousness (orderly, responsible, dependable)
- Neuroticism (calm, not neurotic, not easily upset)
- Openness to experience (intellectual, imaginative, independent-minded)

To evaluate the Big Five, we used the Japanese version of the 10-item personality inventor questionnaire [53].

To analyze the correlation between the Big Five personality factors and the questionnaire items, we conducted a correlation analysis using the Pearson correlation. Table 4.3 shows the correlation analysis results. We found moderate negative correlations between conscientiousness and the conversation's load feeling when using the implicit alignment strategy, and between openness to experience, the conversation's natural feeling, and the overall impressions when using the explicit request strategy. These results indicate that interlocutors with a high rate of conscientiousness tend to rate the mental load feeling of conversations as low, and those with a high rate of openness to experience tend to rate the conversation's natural feeling and overall impression as low. These results follow the overall tendency of the impressions: implicit alignment has a better impression than an explicit request. However, for people with a high conscientiousness, the robot's conversation strategy limits the people's behavior and an explicit request might impose a burden; it would thus be better to avoid using the explicit request strategy. In addition, for people whose openness to experience is high, the conversations with explicit request might be seen as unnatural and the overall

Table 4.3 Correlation between personality and conversation impressions (Pearson's r).

	Load feeling		Troublesome feeling		Natural feeling		Overall impression	
	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit
Extraversion	0.03	0.26	−0.12	0.00	−0.15	−0.39	−0.14	−0.22
Agreeableness	0.08	−0.21	0.29	0.32	0.07	0.11	0.17	0.10
Conscientiousness	−0.07	−0.52**	−0.11	−0.22	−0.19	0.05	−0.15	0.16
Neuroticism	0.10	0.01	−0.02	0.15	−0.08	0.03	0.00	−0.10
Openness to experience	0.03	−0.06	0.14	0.24	−0.41*	−0.04	−0.41*	−0.14

*: $p < .05$ **: $p < .01$

impression might be seen as low. Hence, the explicit request strategy is not suitable for a service robot, which places importance on the impressions of the conversations and the success rate of the tasks.

To behave in accordance with a human's personality, the robot needs to know the personality of the interlocutor beforehand. However, in cases where robots serve many unspecified people in a real environment, it is difficult to know an interlocutor's personality as prior information. If the robot can estimate an interlocutor's personality through a conversation, the robot can behave based on the estimated personality. Past research has reported the relationship between the human features of sound and linguistics, body motion, and personality [54, 55, 56, 57]. By exploiting these phenomena to robots, if the robot estimates an interlocutor's personality, the robots could behave by considering the personality, even in a public environment.

4.5.3 Case of Recognition Failure and its Effects on the Results

We developed a system for indicated object recognition, which uses a combination of existing technology types. If the recognition failed frequently in easy situations for

recognition, the system performance might affect the experimental results. Therefore, we analyzed the conversation with failed recognition to establish what caused the failure. In this experiment, object reference conversations with explicit requests and implicit alignment were conducted 260 times. In the 260 conversations with explicit requests, recognition failed 42 times, and in the conversations with implicit requests, recognition failed 30 times. In four of the failure cases with explicit requests and two with implicit alignment, the interlocutors referred to an object in a way that could not identify an object. For example, the interlocutors referred to an object with speech using a combination of object attributes not existing in the environment, and with speech that indicated multiple objects in the environment without a pointing gesture. In the remaining 38 conversations with an explicit request and 26 conversations with implicit alignment, the humans' references included sufficient information to uniquely identify an object but the system failed to recognize the indicated object. The causes of failures are described as follows. The failure to segment the reference speech section due to intervals and faltering during the speech occurred 14 times with explicit requests and eight times with implicit alignment. Speech using only spacial expressions that the robot cannot recognize such as "front" and "back" occurred zero times with explicit requests and three times with implicit alignment. Partial failure to recognize the object attributes due to the failure of the speech recognition occurred 19 times with explicit requests and eight times with implicit alignment. Failure caused by the inaccuracy of the pointing gesture recognition occurred four times with explicit requests and seven times with implicit alignment. The number of the failures did not differ between the two strategies, although the object arrangement and its combination differed in each session, making difficult to compare the causes and number of failures between the two conversation strategies.

In addition, the recognition performance with implicit alignment described in Subsection 3.4.1 and same objects arrangement as this experiment, one place, is 82%, and big difference is not evident between the results and our experimental results 88%, described in Subsection 4.4.1. Similarly, in the experiment described in Chapter 3, the recognition performance of the conversation that the robot confirms the object in the same way as the explicit request level in this experiment without explicit request in the *Ask* part is 72% (Figure 3.13). This recognition performance of the experiment shown in Figure 3.13 is somewhat lower than the rate of the explicit request level of 84% in this experiment, which suggests that robot's explicit request changed the human's reference behaviors and improved the recognition performance. For these reasons, the system used in this experiment has the same level of performance as systems presented in other studies, and the effects of the system performance on the experimental results are small.

4.5.4 Comparison of Number of Object Attributes and Pointing Gestures in Humans' References

To investigate whether the number of object attributes and pointing gestures in the references changed depending on the robot's conversation strategy, we measured the mean number of object attributes in the speech and pointing gestures for each session in the same way as described in Subsection 3.4.2. Table 4.4 shows the results. The results of a paired t-test showed that no significant difference existed in the mean number of object attributes ($t(25) = 1.677$, $p = .106$, $d = .29$), but a significant difference was found for the applied strategy factor ($t(25) = 4.477$, $p < .001$, $d = .19$), with the explicit request eliciting more pointing gestures than the implicit alignment. These results indicate that humans refer to an object by using a similar amount of object

Table 4.4 Mean number of object attributes and pointing gestures included in references with *SE*.

	Explicit	Implicit
Attributes	2.6 (0.11)	2.5 (0.12)
Pointing gestures	0.98 (0.016)	0.58 (0.093)

information in speech with both the explicit request and the implicit alignment, and they refer to an object with more pointing gestures in the conversations with the explicit request than with the implicit alignment.

For this result and the success rate of the recognition shown in Subsection 4.4.2, it seems that both the explicit request and the implicit alignment strategies elicit sufficient speech and pointing gestures to identify an object with the same level of accuracy. Even in the conversations with the implicit alignment elicits enough pointing gestures to identify an indicated object, and enough object attributes in speech are obtained in each strategy. Therefore, a significant difference of the mean number of pointing gestures exists, and from the viewpoint of the success rate of the recognition, it is considered that there are few substantive differences in the human's reference behaviors. However, in the situation where pointing gestures are important to recognize an indicated object, e.g., robots have no speech recognition function and objects that have similar characteristics are arranged separately, if the robots elicit many pointing gestures by making an explicit request, the recognition performance might improve.

4.5.5 Fairness of Experimental Conditions

In the explicit request condition, the robot requests an interlocutor to refer to an object with all information that is useful for identifying the object. Here, if the

robot requests unnecessary information, such a request might affect the impressions of the conversations. For example, considering the books in our experiment, if each arranged object has the same attribute type, e.g., all books are red colored, all books are emblazoned with a triangle, and Q is printed on all the book covers, a requests by the robot for one more piece of information would be unnecessary because when one attribute of the books match applies, the accordant attribute is not useful for identifying an object and is assumed unnecessary. Therefore, we measured the number of such cases where one attribute of five objects is the same in the conversations with an explicit request. Among the 26 experiment participants, in the experiment of one female, the symbols attribute was assorted, and a triangle was printed on all five books. As this experiment with the female could affect our results, we re-analyzed the results of Section 4.4, excluding this female. The results followed the same tendency of the results for all participants. In the results with an explicit request, for recognition performance, $M = .836$, $SE = .0408$; for load feeling, $M = 3.88$, $SE = .362$ (seven is the most positive); for troublesome feeling, $M = 3.96$, $SE = .324$ (seven is the most positive); for natural feeling $M = 3.32$, $SE = .340$; and for overall impression, $M = 3.72$, $SE = .319$. In the results with the implicit alignment, for recognition performance, $M = .880$, $SE = .0258$; for load feeling $M = 4.84$, $SE = .309$ (seven is most positive); for troublesome feeling, $M = 5.28$, $SE = .268$ (seven is most positive); for natural feeling, $M = 4.08$, $SE = .264$; and for overall impression, $M = 4.20$, $SE = .289$. The results of a paired t-test showed that no significant difference existed for the recognition performance between the explicit request and the implicit alignment ($t(24) = -.967$, $p = .343$, $d = .26$). However, a significant difference existed for all items of the impression of conversations: for load feeling ($t(24) = 2.340$, $p = .028$, $d = .57$); for troublesome feeling ($t(24) = 4.423$, $p < .001$, $d = .89$); for natural feeling ($t(24) = -2.282$, $p = .032$, $d = .50$), and for overall

impression ($t(24) = -2.071$, $p = .049$, $d = .32$). Therefore, although the robot requested information unnecessarily in the conversation of one participant, the effects of the unnecessary request on the results would be small.

In the explicit request, the robot confirmed an object with all necessary information for the recognition. Here, if objects have a huge variety of attributes; for example, an object has more than 10, the robot's confirmation using all the various information might decrease the impression of the conversations. In our experimental settings, objects have three types of attributes, and the robot's confirmation would not be unnatural. Our experiment gathered free description feedback about the robot after the experiment, but no feedback about the robot's confirmation behaviors.

In addition, in the implicit alignment, the robot confirms an object with minimum information to identify the object uniquely, and the way of confirmation is the alignment strategy to reduce ambiguity of human references. Therefore, if the robot could confirm an object with the minimum information in addition to the explicit request, the obtained results are affected by the explicit request and the implicit alignment; the effects of two strategies could not be separated and the combination of the implicit alignment and the explicit request was thus considered invalid.

4.5.6 Limitations

We conducted this experiment in a limited situation, meaning that the participants referred to objects with only three features: color, a geometric symbol, and a letter. In real environments, the features of objects are not limited to three features and thus the variety of the features of objects influence the references. However, since the interaction between a robot and an interlocutor does not depend on features, our findings can be generalized to other objects.

4.5.7 Conclusion

This chapter focused on two interactive strategies for object recognition contexts in conversations with people: explicit request and implicit alignment. We developed a system that recognizes the indicated objects by integrating the speech, pointing, and face direction recognition results, and we experimentally compared the performance and feeling perspectives between the two interactive strategies.

The experimental results indicated that the participants perceived the conversations with the implicit alignment strategy to have a lower mental load and to be less troublesome and more natural than the explicit request strategy. The overall impression of the conversations with the implicit alignment strategy exceeded that of the explicit request strategy. The object reference recognition performance did not differ between the two strategies, indicating that the implicit alignment strategy is better than the explicit request strategy for object recognition contexts in conversations with people. We believe that our findings are useful for the design interaction of social robots that frequently interact with people.

Chapter 5

Gender Differences in Lexical Alignment in Hunan–Robot Interaction

Alignment has often been researched in terms of gender differences. Some past research works on human–human interaction have observed differences in the degree of alignment by gender [58, 59, 60]. Namy *et al.* [59] reported that females aligned with each other in relation to word pronunciation more than did males, and Levitan *et al.* [58] found that alignment in acoustic and/or prosodic features was most prevalent for female–male conversation pairs.

In the alignment between humans and artificial media or robots, other works cited gender differences [31, 36, 61]. Thomason *et al.* [61] found that males aligned more than females in terms of vocal loudness features, and Strupka *et al.* [36] reported that even though humans aligned with robot voices in relation to acoustic energy level, gender had no effect on their voice adjustments. Past research works mentioned gender differences in the degree of alignment [58, 59, 60, 62, 31, 36, 61]. However, since no study has addressed the gender-based differences in lexical alignment in human–robot interaction, we examine these differences, which is worth investigating for the following two reasons. First, investigating gender differences is important for understanding human activity; many researchers have investigated gender differences in various psychological attributes [63, 64]. Hyde [64] gave the following reason for the importance of research on gender differences and similarities in his review article: "stereotypes about psychological gender differences abound, influencing people's behavior, and it

is important to evaluate whether they are accurate". While gender differences have also been investigated in the research fields of alignment, gender differences in lexical alignment between people and robots have not been well investigated. Interaction between people and robots is the new interaction style compared to the interaction between people, and the effect of gender differences in human–robot interaction is different from that in human–human interaction. Revealing gender-based differences in lexical alignment that occur in human–robot interaction is important because identifying the gender differences in lexical alignment between humans and robots would help in the design of human–robot interaction.

This chapter therefore investigates whether gender-based differences in lexical alignment occur in human–robot interactions and discusses a robot's interaction strategies based on gender differences in lexical alignment. We conducted an experiment using a robotic system that interacted with humans in situations where a human referred to an object in an environment and a robot confirmed that indicated object (Figure 3.1).

5.1 Alignment and Gender differences

5.1.1 Gender Differences in Alignment between Humans

Past research has observed differences in the degree of alignment with other humans based on gender [58, 59, 60, 62]. Namy *et al.* [59] gave a group of male and female participants a single-word shadowing task to investigate gender differences in vocal alignment. Their experimental results suggest that female shadowers are more likely to align with female speeches than male shadowers. Levitan *et al.* [58] measured alignment in three acoustic or prosodic features (intensity, pitch, and jitter) that were extracted

from the speech of participants playing a cooperative computer game and found that alignment is most prevalent among female–male pairs, followed by female–female pairs. The alignment of the male–male pairs was the lowest. Pardo [60] investigated the alignment of pronunciation in task-oriented conversations and found that, overall, male talkers aligned with each other more than did females.

Gender differences in the degree of alignment between humans has also been observed in past work on alignment related to vocal interaction between humans. However, to the best of our knowledge, no research has investigated lexical alignment. Furthermore, the gender differences in the degree of alignment reported by past research is inconsistent. For example, while Namy *et al.*[59] argued that females are more likely to align with each other than are males, Pardo's results showed that males are more likely to align with each other than are females [60].

5.1.2 Gender Differences in Alignment between Humans and Artificial Media or Robots

Gender differences in the degree of alignment have also been mentioned by past research examining human–artificial media or robot interaction [31, 36, 61]. Thomason *et al.* [61] investigated the relationships between acoustic or prosodic alignment to a tutoring dialogue system and concluded that males aligned more than females with loudness features. Strupka *et al.* [36] investigated acoustic or prosodic alignment in human–robot dialogues. Their results showed that the gender of the robot's voice marginally affected the acoustic or prosodic alignment, but they found no effect of human gender. Iio *et al.* [31] experimented with a remotely operated robot and investigated whether human pointing aligned with the robot's gestures. Their analysis concluded that the human gender differences had no effect on the alignment

of pointing gestures.

The field of human–artificial media or robot interaction of alignment has also examined the effect of gender differences on the degree of alignment. However, since no past research has—to the best of our knowledge—treated lexical alignment, we examine whether gender differences affect the degrees of lexical alignment.

5.2 Interaction Design

To investigate the effect of gender differences on lexical alignment between humans and robots, we used object reference conversations (Figure 3.1), which focus on confirmation behavior that is often observed in human–human communication. If a person cannot confidently understand which object was being referenced, she is likely to ask for confirmation. Furthermore, people sometimes confirm the referenced object, even if the object being referenced object is clear, to avoid discrepancies in the interpretation. Such conversations are already being used in human–robot interaction research fields to explore lexical alignment as well as the alignment of pointing gestures in human–robot interaction [31, 21, 37, 65, 38].

Object reference conversations comprise four parts: *Ask*, *Refer*, *Confirm*, and *Answer*. First, a robot asks an interlocutor to refer to an object in an environment (*Ask*). Next, the interlocutor refers to an object (*Refer*), and the robot confirms the object to which the interlocutor referred (*Confirm*). Then the interlocutor answers whether the object confirmed by the robot is correct (*Answer*).

Based on past research works, which investigated whether lexical alignment occurs in human–robot interaction and to what degree [38, 65], we employ two interaction strategies for the robot. Both strategies are multi-modal interactions considering speech and gestures. We used the multi-modal strategies rather than

strategies that use specific modalities because human–human interaction is multi-modal, and multi-modal interaction is needed for natural interactions between people and robots. The use of only partial modality is unnatural for interactions with people. Multi-modal interaction is particularly important for lexical alignment because a human's gestures are reported to affect lexical alignment. Holler and Wilkin [43] reported that lexical alignment becomes suppressed when humans align with their interlocutor's gestures. Iio *et al.* [30] also suggested that lexical alignment about an object's attributes becomes suppressed when humans align with a robot's pointing gesture.

Therefore, we did not investigate the differences of each modality (e.g., gesture only) or a different interaction style (e.g., a typed response on a keyboard) partially; rather, we were interested in the gender differences under a human-like conversation style because this style would be common for social robots that act in real environments. To investigate the gender differences in the human–robot interaction research field, we focused on the two major conversation strategies in human–robot interaction under object-reference conversations: explicit and implicit. These conversation strategies are already used to investigate the degree of alignment in human–robot interaction [37, 38, 65]; therefore, using these two strategies would be appropriate for our purpose. The details of each strategy are described in the following Subsection 5.2.1 and 5.2.2.

5.2.1 Implicit Alignment Strategy

One approach is the implicit alignment strategy proposed by Kimoto *et al.* [37, 38]. In this strategy, a robot makes confirmations that contain minimum information for distinguishing objects. Figure 4.2 shows an example of an object reference conversation

with an implicit alignment strategy.

This strategy exploits alignment in object reference conversations based on three alignment phenomena—lexical alignment, gestural alignment, and alignment inhibition. A robot should use minimum information for distinguishing among objects in the environment. Alignment inhibition is a formation phenomenon that decreases in some conversations. Through this design, people learn to make references that include sufficient information to identify the objects by reducing the alignment inhibitions. Kimoto *et al.* [37, 38] implemented this design in the *Confirm* part of the object reference conversation.

5.2.2 Explicit Request Strategy

Another approach is the explicit request strategy in which a robot asks the interlocutors to make a reference that includes as much information as possible and requests that they use the information that was missing from their previous references. Figure 4.1 shows an example of an object reference conversation with an explicit request strategy.

This strategy is based on the following considerations. If an interlocutor refers to an object, as prompted by the robot, it will likely recognize it with a high performance. If the interlocutor fails to follow the robot's requests, the robot should also request that the interlocutor use all of the instructed information for the following object references. This suggestion encourages the interlocutor to obey the robot's requests in subsequent conversations.

5.3 System

We developed a system based on past works that implemented implicit alignment and/or explicit requests for object reference conversations [37, 38, 65]. The system

comprises four parts: sensors, an indicated object recognition function, an object information database, and a conversation strategy selection function. When a user refers to an object, the indicated object recognition function identifies the user's reference behavior and estimates the indicated object. The conversation strategy selection function chooses a robot behavior that corresponds to the implemented strategy and sends a behavior command to the robot, which confirms the indicated object and asks a user to refer to it in the next conversation in a manner decided by the conversation strategy selection. Figure 4.3 shows the architecture of our developed system.

The system can also have object reference conversations as a basic function. In its *Ask* and *Confirm* parts, the robot performs a behavior that corresponds to whichever strategy was used by the robot.

5.3.1 Robot

In this study, we used Robovie-R ver.2, a humanoid robot developed by the Intelligent Robotics and Communication Labs, ATR, which has a human-like upper body designed for communication with humans. The robot has three DOFs for its neck and four for each arm, and its body has an expressive ability for object reference conversations. We used XIMERA for speech synthesis [44]. The robot is is 1100 mm tall, 560 mm wide, 500 mm deep, and weighs about 57 kg.

5.3.2 Indicated Object Recognition Function

To develop this function, we implemented an algorithm [37, 38, 65] that combines the speech recognition, pointing gesture recognition, and face direction recognition results.

5.3.2.1 Speech Recognition Module

The speech recognition module receives human speech, which refers to an object and outputs the normalized reference likelihood of each object based on speech recognition. To calculate the likelihood, we used the number of attributes in human speech [37, 38], which was captured using a microphone attached to a human's collar. In this system, we used a speech recognition engine called Julius, which gives a good performance in Japanese [45].

5.3.2.2 Pointing Gesture Recognition Module

The pointing gesture recognition module obtains the body frame data from a depth image sensor called Kinect for Windows v2 and outputs the normalized reference likelihood of each object based on pointing gesture recognition. We modeled the likelihood as the difference from the pointing vector (between the human head and the tip of the human hand) to a vector between the human head and an object with a normal distribution function $N(0, 1)$.

5.3.2.3 Face Direction Recognition Module

The face direction recognition module obtains the face direction vector from the depth image sensor and outputs the reference likelihood based on the face direction recognition. We modeled the likelihood based on an angle parallel to the plane of the floor between the face direction vector and a vector between a human head and an object. If the vector is less than $11\pi/18$ rad, the person is considered to be viewing the object because a human's field of view is $11\pi/18$ rad at most [46, 47]; its likelihood is 1, and otherwise 0. The likelihoods are finally normalized from 0 to 1.

5.3.2.4 Integration Module

The integration module merges the reference likelihoods of the speech and both the pointing gesture and face direction recognitions. These three likelihoods are summed and normalized. The object with the highest likelihood is estimated to be the one indicated by the interlocutor.

5.3.3 Conversation Strategy Selection Function

The conversation strategy selection function determines how the robot confirms the indicated object (*Confirm* behavior) and how it asks an interlocutor to refer to it (*Ask* behavior) in subsequent conversations. The conversation contents of the *Confirm* and *Ask* behaviors reflect whether the implicit alignment strategy or the explicit request strategy is used.

When using the implicit alignment strategy, this function chooses the *Confirm* behavior and adopts the implicit alignment strategy, and the robot confirms the indicated object with minimum information for distinguishing among objects. The *Ask* behavior does not adopt a particular strategy, and the robot does not explicitly instruct the participants how to make references.

When using the explicit request strategy, this function chooses the *Ask* behavior and adopts the explicit request approach, and the robot explicitly provides requests about how to refer to objects. The *Confirm* behavior does not adopt a particular strategy, and the robot confirms the indicated object by pointing and verifying all of the information about it.

5.4 Experiment

5.4.1 Hypotheses and Predictions

Some past research works reported a gender effect on alignment. However, such gender differences reported by past research are inconsistent. One reported tendency is that females align with each other more than do males [58, 59], but another argues the opposite [60, 61]. Because predicting the effects of gender differences on alignment is difficult, we made two contradictory hypotheses about the effects of gender differences on lexical alignment in human–robot interaction.

Hypothesis about female-dominant differences in lexical alignment in human–robot interaction

Past research identified female-dominant differences in alignment. Namy *et al.* [59] reported that females are more likely to align with robots than are males even when social interaction is severely limited. Their participants performed a shadowing task in which they sat alone in a room and repeated single words uttered by various speakers over headphones. Levitan *et al.* [58] measured alignment on acoustic or prosodic features by analyzing the speech of participants who were playing cooperative computer games for female–female, female–male, and male–male dyads and found that alignment is most prevalent for female–male pairs, followed by female–female pairs. Male–male pairs aligned the least. Although their results did not show that female–female pairs aligned the most, the least aligned pairs were the male–male pairs, thus suggesting that females align with each other more than do males. We therefore believe that females will align more with interlocutors than males. Based on these considerations, we made the following predictions:

Prediction 1-a: Females will lexically align more with a robot interlocutor than will

males.

Hypothesis about male-dominant differences in lexical alignment in human–robot interaction

Past research has identified male-dominant differences in alignment. Pardo [60] concluded that the speech of talkers became more similar to the pronunciation of their partner's speech during conversational interactions. She reported that overall, male talkers were more aligned with each other than female talkers. Thomason *et al.* [61] investigated whether students acoustically and/or prosodically aligned with a tutoring dialogue system. Each student verbally responded to either pre-recorded or synthesized tutor questions. Their results suggested that males were significantly more aligned than females to minimum and maximum features of loudness. Therefore, we believe that males will align more with interlocutors than females. Based on these considerations, we made the following hypothesis:

Prediction 1-b: Males will lexically align more with a robot interlocutor than will females.

5.4.2 Environment

Our participants sat in front of the robot. We arranged books as objects in the environment by following past works that used object reference conversations [37, 38, 65]. Five books were placed in a 1.5 m by 3.3 m rectangular area between the robot and the participant and grouped closely together without overlapping, approximately 0.6–2.6 m from the participants. Figure 5.1 shows the experimental environment.

We controlled the attributes of the books based on past research work that focused on object reference conversations [37, 38, 65]. All the books were 21 cm by 27.5

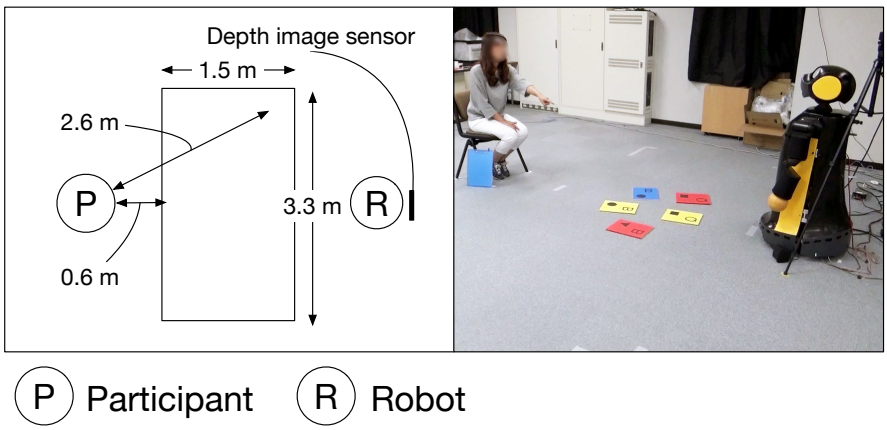

Figure 5.1 Experimental environment.

cm, and their attributes were a color (red, blue, or yellow), a symbol (a circle, a triangle, or a square), and a letter (Q or B) on the cover. We prepared 18 books to satisfy all combinations of attributes.

5.4.3 Conditions

We controlled the strategy that was applied to our developed system (applied strategy factor) using two levels: implicit alignment and explicit requests. These levels were respectively applied to the *Confirm* and *Ask* parts of the object reference conversations. The applied strategy factor had a within-participant condition. There was no difference in the manner of recognizing the interlocutor's reference behavior or estimating the indicated object.

5.4.3.1 Implicit Alignment Condition

In the implicit alignment condition, unlike the explicit request condition, the robot did not explicitly provide requests about the reference style; it just said, "Please choose a book" in the *Ask* part.

However, in the *Confirm* part, the robot said a different sentence. For this purpose, we implemented an implicit alignment design for the reference behavior. In this condition, the robot confirmed the object with minimum information for distinguishing among the objects; confirmations were based on the implicit alignment strategy of references. This approach determined the robot's object reference behaviors, i.e., with or without a pointing behavior and speech contents, by considering the objects' position relationships and characteristics. The robot pointed to the object to reduce the number of possible candidates for the referenced object. The speech format of the confirmations is the sequence of object attributes. For example, the robot asks, "That yellow book with a triangle on its cover?" or "That blue book?"

5.4.3.2 Explicit Request Condition

In the explicit request condition, the robot asks interlocutors to make a reference that includes as much information as possible in the *Ask* part.

The speech format of the explicit requests includes two sentences. The first is used every time, and the second is used only when a participant fails to use all of the information requested in the first sentence in the previous reference.

For example, the robot asks, "Can you refer to the book by its color and the symbol and letter on its cover as well as by pointing and looking at it? Please refer to its letter and point at it."

In the *Confirm* part of this condition, since the robot verified the objects with all of the information, it gave every attribute of an object and pointed during the confirmations.

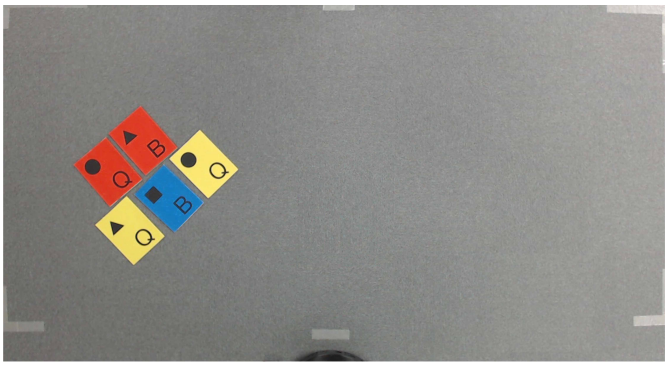

Figure 5.2 Example of book arrangements.

5.4.4 Procedure

We conducted our experiment as follows. First, we explained the experiment to the participants and asked them to sign consent forms. Next, we gave them the following instructions verbally: "The robot can recognize human speech, pointing gestures, and face directions. It will ask you to indicate a book. Do so as if you were addressing a person."

After the requests, the participants selected five books from the 18 and arranged them based on the experimenter's request: "Please arrange the books in one place." Figure 5.2 shows an example of the arrangement. After that, the participant repeated the object reference conversations 10 times. We call this set of 10 object references the conversation sessions, which were conducted using both applied approach conditions: explicit request and implicit alignment. The participants answered questionnaires about their impressions of the conversations after each conversation session. We counterbalanced the order of the interactive strategy conditions.

5.4.5 Measurement

5.4.5.1 Information Amount of Reference

To investigate the change of reference styles, we measured the mean number of object attributes (color, symbol, and letter) used in the participant references in each session.

According to lexical alignment findings between humans and robots, humans tend to use the same word as the robots in conversations [30]. This finding suggests that if a robot uses the word "blue" as a color attribute, humans will avoid the word "cyan" and use blue instead. Therefore, we measured the mean number of object attributes contained in the robot's object information database and used in the *Confirm* part.

5.4.5.2 Reference Redundancy of Speech

The reference redundancy of speech is defined as the difference between the number of object attributes in the participant's references and the minimum number of attributes used for uniquely identifying the referenced objects in the environment in each session. Our objects have three attributes (color, symbol, and letter), and the number of object attributes in the participant's references was defined as 0 to 3. For example, if a participant's reference has no attributes, the numbers of object attributes is 0. If a participant's reference has all three attributes (color, symbol, and letter), the number of object attributes is 3. The minimum number of attributes to uniquely identify the indicated object in the environment ranges from 1 to 3. Therefore, the reference redundancy ranges from -3 to 2. For example, if a participant refers to a book with no attributes (i.e., "that book" or "this book") in the environment where all the attributes are needed to uniquely identify the object (the minimum number of attributes is 3), the reference redundancy of speech is -3 because none of the required attributes are mentioned. We measured the reference redundancy of speech for the following two

reasons.

First, lexical alignment not only increases the use of the word contained in the robot's attributes, but it also leads to alignment of word selection or combination. For example, if a robot uses "blue and B" when it refers to an object, humans tend to use the same combination of words (color and symbol). Such alignment, called word selection or combination, is also observed in human–robot interaction [30, 15]. In our experiment, the robot selects words based on two strategies: implicit alignment and explicit request. As mentioned in Subsection 5.4.3, in both strategies, the robot uses a word combination that can uniquely identify the objects referenced in the environment. With such lexical alignment, humans tend to use words that can identify the objects. For example, in an environment that only has red books, humans rarely use "That red book" as the reference, but instead they say "That red book with a circle and a B on its cover" because the second reference method clearly identifies the object in the environment.

Second, objects in an environment differ with respect to conversation sessions and participants, and the value of the object attributes in a participant's references depends on the environment. For example, the value of "red" as a color attribute in an environment that only has red books is much lower than its value in an environment that has red, blue, and yellow books.

5.4.6 Participants

Twenty people (10 females and 10 males with an average age of 35.5 years, $SD = 9.9$) participated in our experiment. The number of subjects was determined based on past research works about alignment. Five of the seven past works we cited in Section 5.1 investigated the effects of gender differences on alignment using fewer than 20 subjects:

Levitan *et al.* [58], Namy *et al.* [59], Pardo [60], Iio *et al.* [31] and Strupka *et al.* [36]. For example, in the Namy *et al.*'s shadowing task, eight female and eight male shadowers repeated words sounded from headphones [58], and in Iio *et al.*'s task, 10 females and eight males participated in conversations with the robot [31]. Although the experimental procedures of the five past research works are all different and thus comparing the number of subjects is difficult, the numbers of female and male subjects is 10 or less than 10 of each [58, 59, 60, 31, 36].

5.5 Results: Verification of Prediction 1

Figure 5.3 shows the results of the information amount of the references. We conducted a two-factor mixed ANOVA for both applied strategy and gender factors, and we identified the significant main effects in the applied strategy factor ($F(1, 18) = 6.616$, $p = .019$, $\eta_p{}^2 = .269$). We found no significance in the gender factor ($F(1, 18) = 2.646$, $p = .121$, $\eta_p{}^2 = .128$) and no significant interaction ($F(1, 18) = .952$, $p = .342$, $\eta_p{}^2 = .050$). These results showed that the number of object attributes in the references with explicit request was significantly larger than the number of object attributes in references with implicit alignment. However, these results showed no gender-based differences in the information amount of the references.

Figure 5.4 shows the results of the reference redundancy of speech. We conducted a two-factor mixed ANOVA for both applied strategy and gender factors and found significant main effects in the applied strategy factor ($F(1, 18) = 4.485$, $p = .048$, $\eta_p{}^2 = .199$) and gender factor ($F(1, 18) = 4.423$, $p = .050$, $\eta_p{}^2 = .197$); we found no significant interaction ($F(1, 18) = 2.382$, $p = .140$, $\eta_p{}^2 = .117$). These results showed that reference redundancy with an explicit request was significantly larger than the reference redundancy with an implicit alignment. The results also showed that the

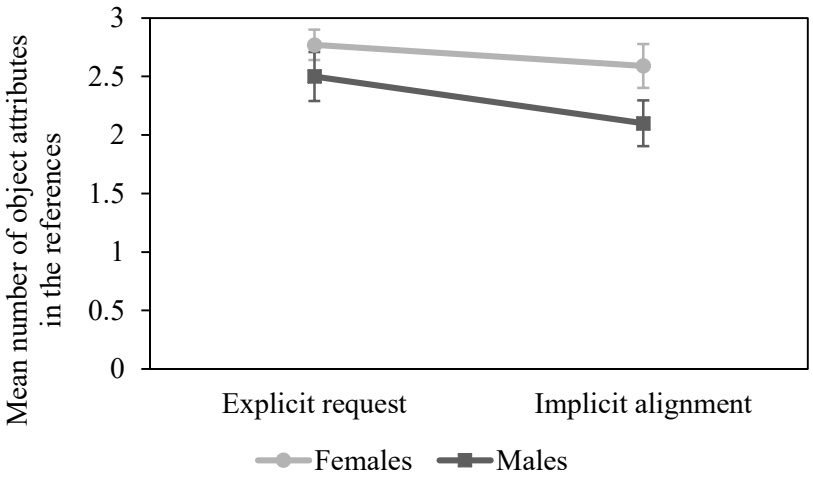

Figure 5.3 Information amount of references with *SE*.

reference redundancy of females' speech exceeds that of males.

These results on the amount of information about references and the reference redundancy of speech partially support prediction 1-a but do not support prediction 1-b.

5.6 Discussion

5.6.1 Implication

Our experimental results showed that females refer to objects with references that have a higher reference redundancy of speech than that of males. The reference redundancy of speech reflects how useful the references are for identifying objects, and our experimental results suggest that robots need to change their interaction strategies for effective alignment with human references to useful references to identify objects in object reference conversations. For example, when a robot uses the implicit alignment strategy, the reference redundancy of male speech is relatively lower than that of

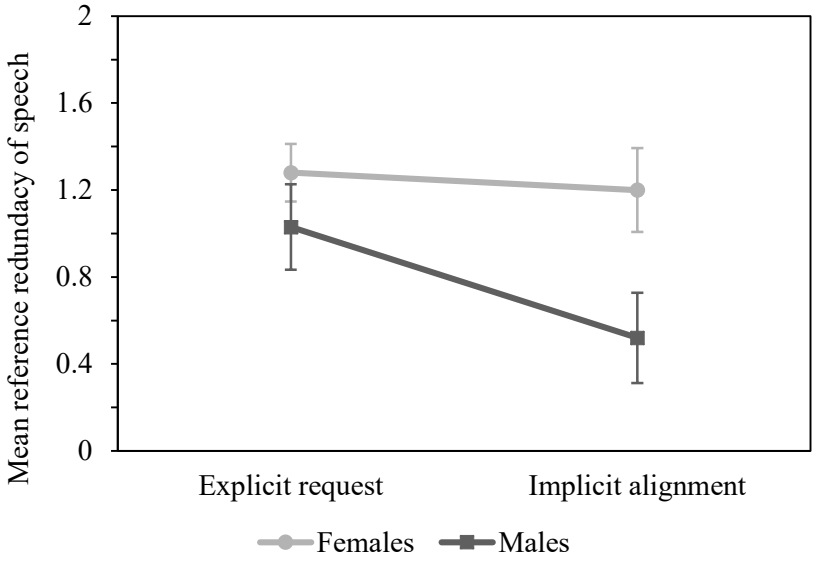

Figure 5.4 Reference redundancy of speech with *SE*.

females; therefore, robots should choose the explicit request strategy to obtain useful information to identify objects from males. Since the overall conversation impressions did not differ by gender or applied strategy factors, an explicit request to males by robots would have fewer disadvantages. For these reasons, considering gender effects on lexical alignment is important for designing conversation strategies for robots. Our findings might be integrated in not only object reference conversation contexts but also other conversation contexts, since lexical alignment is not a phenomenon that is only observed in object reference conversations.

Regarding the information amount, there were no gender differences in lexical alignment, although gender differences were evident for the reference redundancy of speech. This result suggests that, even though males aligned with as many words as females, they aligned with fewer word combinations that could be used to uniquely identify the referenced object in the environment than did females. Namy *et al.* [59] found that in the shadowing tasks, females are vocally more likely to align with each

other than are males, and they considered this finding was because females are more sensitive to the indexical features of interlocutors. If sensitivity to conversational features differs by gender, the difference of sensitivity might explain the discrepancy of our results between the information amount and reference redundancy.

We found significant main effects in the gender factor about the reference redundancy of speech ($F(1, 18) = 4.423$, $p = .050$, ${\eta_p}^2 = .197$). Although the interpretation of effect size varies by experiment, Cohen [66] offered standard interpretations of ${\eta_p}^2$ as benchmarks: small, medium, and large effects would be reflected in ${\eta_p}^2 = .0099$, .0588 and .1379, respectively. Compared to Cohen's benchmarks, the gender factor has a large effect on the reference redundancy of speech.

5.6.2 Conversation Impressions

To investigate the participants' impressions of the conversations, we measured a questionnaire item—overall conversation impression of the robot—which we evaluated using a seven-point scale ranging from 1 (disagree) to 7 (agree). Figure 5.5 shows the questionnaire results for the overall conversation impressions. We conducted a two-factor mixed ANOVA for two factors—applied strategy and gender—and found no significance in the applied strategy factor ($F(1, 18) = 2.751$, $p = .115$, ${\eta_p}^2 = .133$), no significance in the gender factor ($F(1, 18) = 1.254$, $p = .278$, ${\eta_p}^2 = .065$), and no significant interaction ($F(1, 18) = .306$, $p = .587$, ${\eta_p}^2 = .017$).

These results show that the overall conversation impressions did not differ based on the applied strategies and genders.

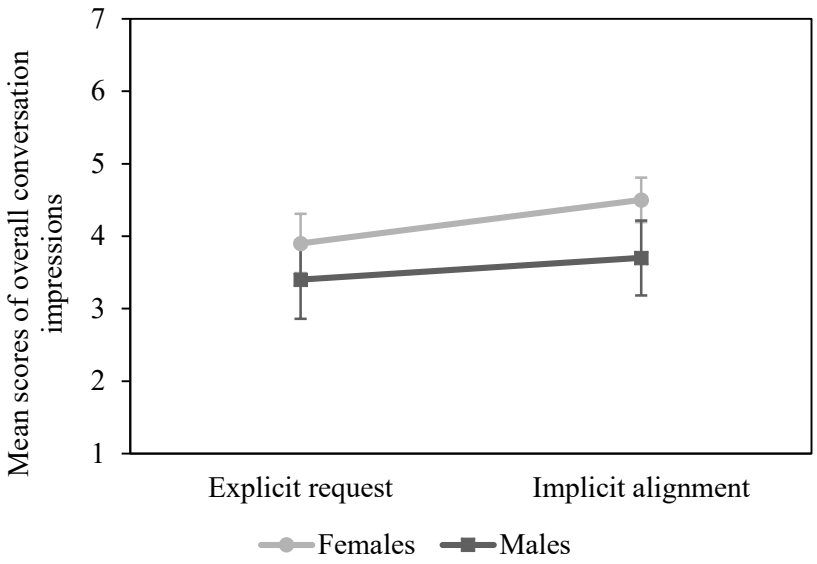

Figure 5.5 Overall conversation impression with *SE*.

5.6.3 Limitations

We conducted our experiment in a limited situation, meaning that the participants referred to objects using only three features: color, a symbol, and a letter. In real environments, the features of objects are not limited and influence the reference methods. However, since the interaction manner between a robot and an interlocutor does not depend on features, our findings can be generalized to other objects.

Since our experiment was conducted with an existing robot named Robovie-R ver.2, robot generality is also limited. Some past research investigated gender differences on lexical alignment from the viewpoint of gender pairs [58, 60, 36]. Robovie-R ver.2 and its synthesized speech have no intended gender. If a robot and/or its speech are designed to represent a specific gender, the robot's gender will also influence lexical alignment in conversations.

5.6.4 Conclusion

In this chapter, we investigated gender differences on lexical alignment in object reference conversation contexts between humans and robots by employing two interaction strategies based on related works: implicit alignment and explicit request. We developed a system that recognized the indicated objects and had object reference conversations with humans. The experimental results indicated that females lexically align more with a robot interlocutor than do males in terms of the reference redundancy of speech. Our female participants aligned more than males and used more references that are useful for uniquely identifying referenced objects in the environment. We believe that our findings of the female-dominant differences in lexical alignment in human–robot interaction will help robotics researchers to design conversation strategies between humans and robots.

Chapter 6
Conclusion

This study proposed a robot conversation strategy to improve the recognition performance of objects when conversing with a person. We considered three phenomena in human–human and human–robot interaction to design the approach: lexical alignment, gestural alignment, and alignment inhibition. Based on these phenomena, we designed robotic behavior policies which suggest that robots should provide minimum information to identify an object and use pointing gestures only if the pointing gestures are useful to identify an object. To verify our design, we developed a robotic system to recognize the object to which people referred and conducted an experiment. The results showed that the proposed approach elicited redundant references from interlocutors and improved the recognition performance of objects to which people referred.

Next, we focused on two interactive strategies for object recognition contexts in conversations with people: explicit request and implicit alignment. We experimentally compared two interactive strategies to determine which approach improves the performance and which approach makes better impressions on people. Even though the results indicated that the participants evaluated the impressions of conversations with the implicit alignment strategy more highly, the recognition performances of the two approaches were not significantly different, indicating that the implicit alignment strategy is better than the explicit request strategy for object reference conversations with people.

Last, we examined the gender differences of lexical alignment and investigated the differences by employing the two interaction strategies: implicit alignment and explicit request. The results indicated that females lexically align more with a robot interlocutor than do males in terms of the reference redundancy of speech. Female participants aligned more with robots than males and used more references that are useful to uniquely identify referenced objects in the environment. Finally, we summarize our findings:

1. The proposed approach implicitly elicits redundant references and improves the performance of the indicated object recognition.

2. Even though the proposed approach forms better impressions than the other interactive approach that explicitly requests clarifications when people refer to objects, the recognition performances of the two approaches are not significantly different.

3. Females lexically align more with robots than do males in terms of the reference redundancy of speech.

Bibliography

[1] M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita, "Interactive humanoid robots for a science museum," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, 2006, pp. 305–312.

[2] H.-M. Gross, H. Boehme, C. Schroeter, S. Mueller, A. Koenig, E. Einhorn, C. Martin, M. Merten, and A. Bley, "Toomas: Interactive shopping guide robots in everyday use - final implementation and experiences from long-term field trials," in *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2009, pp. 2005–2012.

[3] Y. Iwamura, M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita, "Do elderly people prefer a conversational humanoid as a shopping assistant partner in supermarkets?" in *Proceedings of the 6th International Conference on Human-Robot Interaction*, 2011, pp. 449–456.

[4] D. Fischinger, P. Einramhof, K. Papoutsakis, W. Wohlkinger, P. Mayer, P. Panek, S. Hofmann, T. Koertner, A. Weiss, A. Argyros, and M. Vincze, "Hobbit, a care robot supporting independent living at home: First prototype and lessons learned," *Robotics and Autonomous Systems*, vol. 75, pp. 60–78, 2016.

[5] C. J. Calo, N. Hunt-Bull, L. Lewis, and T. Metzler, "Ethical implications of using the paro robot," in *Proceedings of the 2011 AAAI Workshop (WS-2011-2012)*, 2011, pp. 20–24.

[6] S. Ŝabanović, C. C. Bennett, W. Chang, and L. Huber, "Paro robot affects diverse interaction modalities in group sensory therapy for older adults with dementia," in *Proceedings of the 2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*, June 2013, pp. 1–6.

[7] S. Nishio, H. Ishiguro, and N. Hagita, "Geminoid: Teleoperated android of an existing person," in *Humanoid robots: New developments*. InTech, 2007.

[8] H. Ishiguro, "Android science: conscious and subconscious recognition," *Connection Science*, vol. 18, no. 4, pp. 319–332, 2006.

[9] J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, Y. Unno, W. Ko, and J. Tan, "Interactively picking real-world objects with unconstrained spoken

language instructions," in *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 3774–3781.

[10] 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.

[11] 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 *Proceedings of the 2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2008, pp. 241–248.

[12] B. Schauerte and G. A. Fink, "Focusing computational visual attention in multi-modal human-robot interaction," in *Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, 2010, pp. 6:1–6:8. [Online]. Available: http://doi.acm.org/10.1145/1891903.1891912

[13] N. Iwahashi, "A method for the coupling of belief systems through human-robot language interaction," in *Proceedings of the The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003*, Conference Proceedings, pp. 385–390.

[14] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Commun. ACM*, vol. 30, no. 11, pp. 964–971, nov 1987. [Online]. Available: http://doi.acm.org/10.1145/32206.32212

[15] K. Shinozawa, T. Miyashita, M. Kakio, and N. Hagita, "User specification method and humanoid confirmation behavior," in *Proceedings of the 2007 7th IEEE-RAS International Conference on Humanoid Robots*, Nov 2007, pp. 366–370.

[16] E. Wu, Y. Han, D. Whitney, J. Oberlin, J. MacGlashan, and S. Tellex, "Robotic social feedback for object specification," in *Proceedings of the AAAI Fall Symposium on AI for Human-Robot Interaction*, 2015, pp. 150–157.

[17] Y. Kuno, K. Sakata, and Y. Kobayashi, "Object recognition in service robots: Conducting verbal interaction on color and spatial relationship," in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, Sept 2009, pp. 2025–2031.

[18] T. Iio, M. Shiomi, S. Kazuhiko, K. Shimohara, and N. Hagita, "Contribution to performance of object reference recognition by redundancy control of robot speech," *IPSJ Journal*, vol. 53, no. 4, pp. 1251–1268, apr 2012, [published in Japanese].

[19] S. E. Brennan and H. H. Clark, "Conceptual pacts and lexical choice in conversation," *Journal of experimental psychology. Learning, memory, and cognition*, vol. 22, no. 6, pp. 1482–1493, Nov 1996.

[20] S. Garrod and A. Anderson, "Saying what you mean in dialogue: a study in conceptual and semantic co-ordination," *Cognition*, vol. 27, no. 2, pp. 181–218, Nov. 1987.

[21] S. E. Brennan, "Lexical entrainment in spontaneous dialog," in *Proceedings of the International Symposium on Spoken Dialogue*, vol. 96, 1996, pp. 41–44.

[22] H. P. Branigan, M. J. Pickering, and A. A. Cleland, "Syntactic co-ordination in dialogue," *Cognition*, vol. 75, no. 2, pp. B13–B25, 2000.

[23] A. E. Scheflen, "The significance of posture in communication systems," *Psychiatry*, vol. 27, no. 4, pp. 316–331, 1964. [Online]. Available: https://doi.org/10.1080/00332747.1964.11023403

[24] A. Kendon, "Movement coordination in social interaction: Some examples described," *Acta Psychologica*, vol. 32, pp. 101–125, 1970. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0001691870900946

[25] K. Bergmann and S. Kopp, "Gestural alignment in natural dialogue," in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 2012, pp. 1326–1331.

[26] H. Branigan, M. Pickering, J. Pearson, J. McLean, and C. Nass, "Syntactic alignment between computers and people: the role of belief about mental states," in *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, R. Alterman and D. Kirsh, Eds., 2003, pp. 186–191.

[27] J. Gustafson, A. Larsson, R. Carlson, and K. Hellman, "How do system questions influence lexical choices in user answers?" in *Fifth European Conference on Speech Communication and Technology*, 1997.

[28] J. Pearson, J. Hu, H. P. Branigan, M. J. Pickering, and C. I. Nass, "Adaptive language behavior in hci: How expectations and beliefs about a system affect users' word choice," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2006, pp. 1177–1180. [Online]. Available: http://doi.acm.org/10.1145/1124772.1124948

[29] S. Tomko and R. Rosenfeld, "Shaping spoken input in user-initiative systems," in *Eighth International Conference on Spoken Language Processing*, 2004.

[30] T. Iio, M. Shiomi, K. Shinozawa, K. Shimohara, M. Miki, and N. Hagita, "Lexical entrainment in human robot interaction," *International Journal of*

Social Robotics, vol. 7, no. 2, pp. 253–263, Apr 2015. [Online]. Available: https://doi.org/10.1007/s12369-014-0255-x

[31] 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, Nov 2011. [Online]. Available: https://doi.org/10.1007/s12369-011-0112-0

[32] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, 2008, pp. 169–172. [Online]. Available: http://dl.acm.org/citation.cfm?id=1557690.1557737

[33] C.-C. Lee, M. Black, A. Katsamanis, A. C. Lammert, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[34] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[35] 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, how people talk to Robots and Computers. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378216609003282

[36] E. Strupka, O. Niebuhr, and K. Fischer, "Influence of robot gender and speaker gender on prosodic entrainment in hri," 2017.

[37] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, K. Shimohara, and N. Hagita, "Improvement of object reference recognition through human robot alignment," in *Proceedings of the 2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Aug 2015, pp. 337–342.

[38] ——, "Robot confirmation behavior to improve object reference recognition," *Journal of the Robotics Society of Japan*, vol. 35, no. 9, pp. 681–692, 2017, [published in Japanese].

[39] S. E. Brennan, "Conversation with and through computers," *User Modeling and User-Adapted Interaction*, vol. 1, no. 1, pp. 67–86, Mar 1991. [Online]. Available: https://doi.org/10.1007/BF00158952

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

[41] H. Ogawa and T. Watanabe, "InterRobot: a speech driven embodied interaction robot," in *Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000*, Sept 2000, pp. 322–327.

[42] T. Ono, M. Imai, and H. Ishiguro, "A model of embodied communications with gestures between human and robots," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 23, no. 23, 2001.

[43] 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, Jun 2011. [Online]. Available: https://doi.org/10.1007/s10919-011-0105-6

[44] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "Ximera: A new tts from atr based on corpus-based technologies," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[45] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," in *Proceedings of the APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, 2009, pp. 131–137.

[46] I. P. Howard and B. J. Rogers, *Binocular vision and stereopsis*. NY, USA: Oxford University Press, 1995.

[47] T. Hatada, H. Sakata, and H. Kusaka, "Induced effect of direction sensation and display size," *The Journal of the Institute of Television Engineers of Japan*, vol. 33, no. 5, pp. 407–413, 1979, [published in Japanese].

[48] 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, no. 4, pp. 379–402, 2006.

[49] K. K. Ball, V. G. Wadley, and J. D. Edwards, "Advances in technology used to assess and retrain older drivers," *Gerontechnology*, vol. 1, no. 4, 2002.

[50] A. F. SANDERS, "Some aspects of the selective process in the functional visual field," *Ergonomics*, vol. 13, no. 1, pp. 101–117, 1970. [Online]. Available: https://doi.org/10.1080/00140137008931124

[51] Y. Seya and K. Watanabe, "Objective and subjective sizes of the effective visual field during game playing measured by the gaze-contingent window method," *International Journal of Affective Engineering*, vol. 12, no. 1, pp. 11–19, 2013.

[52] O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," *Handbook of personality: Theory and research*, vol. 2, no. 1999, pp. 102–138, 1999.

[53] A. Oshio, S. Abe, and P. Cutrone, "Development, reliability, and validity of the japanese version of ten item personality inventory (tipi-j)," *The Japanese Journal of Personality*, vol. 21, no. 1, pp. 40–52, 2012, [published in Japanese].

[54] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *Proceedings of the 10th International Conference on Multimodal Interfaces*, 2008, pp. 53–60. [Online]. Available: http://doi.acm.org/10.1145/1452392.1452404

[55] L. M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe, "Please, tell me about yourself: Automatic personality assessment using short self-presentations," in *Proceedings of the 13th International Conference on Multimodal Interfaces*, 2011, pp. 255–262. [Online]. Available: http://doi.acm.org/10.1145/2070481.2070528

[56] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, July 2014.

[57] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[58] R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012, pp. 11–19. [Online]. Available: http://dl.acm.org/citation.cfm?id=2382029.2382032

[59] L. L. Namy, L. C. Nygaard, and D. Sauerteig, "Gender differences in vocal accommodation: The role of perception," *Journal of Language and Social Psychology*, vol. 21, no. 4, pp. 422–432, 2002. [Online]. Available: https://doi.org/10.1177/026192702237958

[60] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006. [Online]. Available: https://doi.org/10.1121/1.2178720

[61] J. Thomason, H. V. Nguyen, and D. Litman, "Prosodic entrainment and tutoring dialogue success," in *Artificial Intelligence in Education*, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 750–753.

[62] Z. Xia, R. Levitan, and J. Hirschberg, "Prosodic entrainment in mandarin and english: A cross-linguistic comparison," in *Speech Prosody*, 2014.

[63] J. A. Hall, "Gender effects in decoding nonverbal cues," *Psychological Bulletin*, vol. 85, no. 4, pp. 845–857, 1978.

[64] J. S. Hyde, "Gender similarities and differences," *Annual Review of Psychology*, vol. 65, no. 1, pp. 373–398, 2014. [Online]. Available: https://doi.org/10.1146/annurev-psych-010213-115057

[65] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, K. Shimohara, and N. Hagita, "Alignment approach comparison between implicit and explicit suggestions in object reference conversations," in *Proceedings of the Fourth International Conference on Human Agent Interaction*, 2016, pp. 193–200. [Online]. Available: http://doi.acm.org/10.1145/2974804.2974814

[66] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, 1977.

Publication List

Journal Papers

[1] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, K. Shimohara, and N. Hagita, "Conversation Strategy Comparison between Explicit Request and Implicit Alignment in Object Reference Conversation," *Journal of the Robotics Society of Japan*, vol. 36, no. 6, pp. 441–452, Jul. 2018 [published in Japanese].

[2] M. Kimoto, T. Nakahata, M. Shiomi, T. Iio, I. Tanev, and K. Shimohara, "System Supporting Self-Motivated Video-Viewing Stops for Children," *SICE Journal of Control, Measurement, and System Integration*, vol. 11, no. 1, pp. 48–54, Jan. 2018.

[3] M. Kimoto, T. Iio, M. Shiomi, I. Tanev and, K. Shimohara, "Can Graphical Interaction Increase Feelings of Conveying and Understanding in On-line Group Discussion?," *SICE Journal of Control, Measurement, and System Integration*, vol. 11, no. 1, pp. 55–64, Jan. 2018.

[4] T. Hirano, M. Shiomi, T. Iio, M. Kimoto, I. Tanev, K. Shimohara, and N. Hagita, "How Do Communication Cues Change Impressions of Human-Robot Touch Interaction?," *International Journal of Social Robotics*, vol. 10, no. 1, pp. 21–31, Jan. 2018.

[5] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, K. Shimohara, and N. Hagita, "Gender Effects on Lexical Alignment in Human-Robot Interaction," *IEEJ Transactions on Electronics, Information and Systems*, vol. 137, no. 12, pp. 1625–1632, Dec. 2017.

[6] M. Kimoto and T. Iio and M. Shiomi and I. Tanev, K. Shimohara, and N. Hagita, "Robot Confirmation Behavior to Improve Object Reference Recognition," *Journal of the Robotics Society of Japan*, vol. 35, no. 9, pp. 681–692, Nov. 2017 [published in Japanese].

International Conferences

[1] M. Kimoto, M. Shiomi, T. Iio, K. Shimohara, and N. Hagita, "Calibrating Depth Sensors for Pedestrian Tracking Using a Robot as a Movable and Localized Landmark," in *Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC2018)*, Miyazaki, Japan, Oct. 2018. pp. 345–350.

[2] S. Okumura, M. Kimoto, M. Shiomi, T. iio, K. Shimohara, and N. Hagita, "Do Social Rewards from Robots Enhance Offline Improvements in Motor Skills?," in *Proceedings of the Ninth International Conference on Social Robotics (ICSR2017)*, Tsukuba, Japan, Nov. 2017. pp. 32–41.

[3] Y. Tamura, M. Kimoto, M. Shiomi, T. Iio, K. Shimohara, and N. Hagita, "Effects of a Listener Robot with Children in Storytelling," in *Proceedings of the 5th International Conference on Human Agent Interaction (HAI2017)*, Bielefeld, Germany, Oct. 2017, pp. 35–43.

[4] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, and K. Shimohara, "Relationship between Personality and Robots' Interaction Strategies in Object Reference Conversations," in *Proceedings of the Second International Conference on Electronics and Software Science*, Takamatsu, Japan, Nov. 2016, pp. 128–136.

[5] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, K. Shimohara, and N. Hagita, "Alignment Approach Comparison between Implicit and Explicit Suggestions in Object Reference Conversations," in *Proceedings of the Fourth International Conference on Human Agent Interaction (HAI2016)*, Biopolis, Singapore, Oct. 2016, pp. 193–200.

[6] T. Hirano, M. Shiomi, T. Iio, M. Kimoto, T. Nagashio, I. Tanev, K. Shimohara, and N. Hagita, "Communication Cues in a Human-Robot Touch Interaction," in *Proceedings of the Fourth International Conference on Human Agent Interaction (HAI2016)*, Biopolis, Singapore, Oct. 2016, pp. 201-206.

[7] T. Nagashio, M. Kimoto, M. Shiomi, T. Iio, T. Hirano, I. Tanev, and K. Shimohara, "Supporting Better Sleep by Using a Hugvie," in *Proceedings of the the SICE Annual Conference 2016*, Tsukuba, Japan, Sept. 2016, pp. 834–837. (Abstract)

[8] M. Kimoto, T. Nakahata, T. Hirano, T. Nagashio, M. Shiomi, T. Iio, I. Tanev, and K. Shimohara, "Video Recommendation System that Arranges Video Clips based on Pre-defined Viewing times," in *Proceedings of the 18th International Conference on Human-Computer Interaction (HCII2016)*, Toronto, Canada, July. 2016, pp. 478–486.

[9] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, K. Shimohara, and N. Hagita, "Improvement of Object Reference Recognition through Human Robot Alignment," in *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN2015)*, Kobe, Japan, Aug. 2015, pp. 337–342.

[10] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, and K. Shimohara: "Self-Organizing of Information by Rhizomic-Link which Autonomously Grows: Modeling and Evaluation of Rhizomic-Link Mechanism," in *Proceedings of the 34th Chinese Control*

Conference and SICE Annual Conference 2015, Hangzhou, China, July. 2015, pp. 347–350. (Abstract)

[11] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, and K. Shimohara, "Can Graphical Interaction Affect Mutual Understanding?," in *Proceedings of the International Conference on Electronics and Software Science*, Takamatsu, Japan, July. 2015, pp. 78–86.

[12] K. Kimura, M. Kimoto, and K. Shimohara, "Do!PaPi: Platform for E-book as Media for Relationality," in *Proceedings of the International Conference on Humanized Systems 2013*, Kagawa, Japan, Sept. 2013, pp. 50–53.

Domestic Conferences

[1] M. Kimoto, T. Iio, M. Shiomi, K. Shimohara, and N. Hagita, "Impression of a Caregiving Task by Communicating the Robot's Motion State through Robots Conversation," Symposium on Human Interface 2018, Tsukuba, Sept. 2018, 64P, [published in Japanese].

[2] M. Kimoto, T. Iio, M. Shiomi, K. Shimohara, and N. Hagita, "Calibrating Depth Sensors for Pedestrian Tracking Using a Mobile Robot," the 36th Annual Conference of the RSJ, Aichi, Sept. 2018, 3A1-04, [published in Japanese].

[3] T. Hirano, M. Shiomi, M. Kimoto, T. Iio, K. Shimohara, and N. Hagita, "Gaze-Height and Speech-Timing Effects about Feelings of Robot-Initiated Touch," the 36th Annual Conference of the RSJ, Aichi, Sept. 2018, 2D1-05, [published in Japanese].

[4] A. Saito, M. Kimoto, M. Shiomi, T. Iio, S. Otani, I. Tanev, K. Shimohara, and N. Hagita, "The Influence of Existence of Multiple Robots in Human-Robot Interaction on Human Vocabulary Selection," 45th SICE Symposium on Intelligent Systems, Osaka, Mar. 2018, A3-1, [published in Japanese].

[5] S. Otani, M. Kimoto, M. Shiomi, T. Iio, A. Saito, I. Tanev, K. Shimohara, and N. Hagita, "The Influence of Priming Information for the Robots on Peer Pressure," 45th SICE Symposium on Intelligent Systems, Osaka, Mar. 2018, A3-2, [published in Japanese].

[6] I. Brison, K. Kimura, M. Kimoto, I. Tanev, and K. Shimohara, "Proposal of a Way to Build a Category Based on Semilattice Structure and a Verification of its Usage," 45th SICE Symposium on Intelligent Systems, Osaka, Mar. 2018, A3-4, [published in Japanese].

[7] S. Okumura, M. Kimoto, M. Shiomi, T. Iio, I. Tanev, K. Shimohara, and N. Hagita, "Influence of Social Rewards by Robots on Enhancement of Consolidation in Motor Skill," the 35th Annual Conference of the RSJ, 3F2-01, Saitama, Sept. 2017, [published in Japanese].

[8] Y. Tamura, M. Kimoto, M. Shiomi, T. Iio, I. Tanev, K. Shimohara, and N. Hagita, "Effects of a Listener Robot with Children in Storytelling," the 35th Annual Conference of the RSJ, 3D3-05, Saitama, Sept. 2017, [published in Japanese].

[9] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, K. Shimohara, and N. Hagita, "Analysis of Relationship between Human's Personality Traits and Robots' Interaction Strategies in Object Reference Conversations," at 44th SICE Symposium on Intelligent Systems, Tokyo, Mar. 2017, A4-3, [published in Japanese].

[10] S. Okumura, M. Shiomi, T. Iio, M. Kimoto, I. Tanev, K. Shimohara, and N. Hagita, "Influence of Praise by Multiple Robots on Enhancement of Consolidation in Motor Skill," 44th SICE Symposium on Intelligent Systems, Tokyo, Mar. 2017, A4-2, [published in Japanese].

[11] Y. Tamura, M. Kimoto, M. Shiomi, T. Iio, I. Tanev, K. Shimohara, and N. Hagita, "Research on Effect by the Usage of Multiple Robots in the Story-telling to a Child," 44th SICE Symposium on Intelligent Systems, Tokyo, Mar. 2017, A4-1, [published in Japanese].

[12] A. Noshita, M. Kimoto, K. Kimura, I. Tanev, and K. Shimohara, "Visualization and Support of Thinking Process through Web Browsing," SICE Symposium on Systems and Information 2016, Shiga, Dec. 2016, SS02-9, [published in Japanese].

[13] Y. Tamura, M. Kimoto, M. Shiomi, T. Iio, I. Tanev, and K. Shimohara, "Effects and Usage of Multiple Robots with Children in Storytelling," SICE Symposium on Systems and Information 2016, Shiga, Dec. 2016, SS02-11, [published in Japanese].

[14] S. Okumura, M. Kimoto, M. Shiomi, T. Iio, I. Tanev, and K. Shimohara, "Influence of Praise by Multiple Robots on Enhancement of Consolidation in Motor Skill," SICE Symposium on Systems and Information 2016, Shiga, Dec. 2016, SS02-13, [published in Japanese].

[15] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, K. Shimohara, and N. Hagita, "Comparison of a Performance and Impressions between Explicit Request and Implicit Alignment in Object Reference Conversation," the 34th Annual Conference of the RSJ, Yamagata, Sept. 2016, 3W2-04, [published in Japanese].

[16] T. Hirano, M. Shiomi, T. Iio, M. Kimoto, I. Tanev, K. Shimohara, and N. Hagita, "Gaze Behaviors and Touch Styles Effects about Feelings of Robot-Initiated

Touch," the 34th Annual Conference of the RSJ, Yamagata, Sept. 2016, 3W3-06, [published in Japanese].

[17] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, and K. Shimohara, "Visualizing Processes of Collecting, Editing and Expression through Web Browsing," 43rd SICE Symposium on Intelligent Systems, Muroran, Mar. 2016, A3-1, [published in Japanese].

[18] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, K. Shimohara, and N. Hagita, "Performance Improvement of Indicated Object Recognition through Gestural and Speech Alignment," the 33rd Annual Conference of the RSJ, Tokyo, Sept. 2015, 3G3-07, [published in Japanese].

[19] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, and Katsunori Shimohara, "Self-Organizing of Information by Rhizomic-Link which Autonomously Grows: Modeling and Evaluation of Rhizomic-Link Mechanism," 42th SICE Symposium on Intelligent Systems, Kobe, Mar. 2015, F-08, [published in Japanese].

[20] M. Kimoto, I. Tanev, and K. Shimohara, "Influence of Diversity in Interpretation on Communications," 41th SICE Symposium on Intelligent Systems, Tokyo, Mar. 2014, A21-3, [published in Japanese].

[21] M. Kimoto, I. Tanev, and K. Shimohara, "Influence of Diversity in Interpretation on Communications," SICE Symposium on Systems and Information 2013, Shiga, Nov. 2013, SS6-11, [published in Japanese].

Awards

[1] The Fourth International Conference on Human Agent Interaction (HAI2016), **Best Student Paper Award Candidates**, "Alignment Approach Comparison between Implicit and Explicit Suggestions in Object Reference Conversations," 2016.

[2] The Fourth International Conference on Human Agent Interaction (HAI2016), **Best Student Paper Award Candidates**, "Communication Cues in Human-Robot Touch Interaction," 2016.

Grant

[1] Grant-in-Aid for JSPS Research Fellow, Japan Society for the Promotion of Science (JSPS), Apr. 2018–Mar. 2020.