applied sciences

MDPI

Article

Coordinating Entrainment Phenomena: Robot Conversation Strategy for Object Recognition

Mitsuhiko Kimoto [1,2,*], Takamasa Iio [2,3], Masahiro Shiomi [2] and Katsunori Shimohara [4]

[1] Faculty of Science and Technology, Keio University, Kanagawa 223-8522, Japan
[2] Interaction Science Laboratories, ATR, Kyoto 619-0288, Japan; iio@iit.tsukuba.ac.jp (T.I.); m-shiomi@atr.jp (M.S.)
[3] Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba 305-8573, Japan
[4] Faculty of Science and Engineering, Doshisha University, Kyoto 610-0321, Japan; kshimoha@mail.doshisha.ac.jp
* Correspondence: kimoto@ailab.ics.keio.ac.jp

Abstract: This study proposes 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. Research conducted to improve the performance of indicated object recognition is divided into two main approaches: development and interactive. The development approach addresses the development of new devices or algorithms. Through human–robot interaction, the interactive approach improves the performance by decreasing the variability and the ambiguity of the references. Inspired by the findings of entrainment and entrainment inhibition, this study proposes a robot conversation strategy that utilizes the interactive approach. While entrainment is a phenomenon in which people unconsciously tend to mimic words and/or gestures of their interlocutor, entrainment 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 that elicits clear references. We experimentally compared this strategy with the other interactive strategy in which a robot explicitly requests clarifications when a human refers to an object. We obtained the following findings: (1) The proposed strategy clarifies human references and improves indicated object recognition performance, and (2) the proposed strategy forms better impressions than the other interactive strategy that explicitly requests clarifications when people refer to objects.

Keywords: human–robot interaction; indicated object recognition; entrainment; alignment

check for updates

Citation: Kimoto, M.; Iio, T.; Shiomi, M.; Shimohara, K. Coordinating Entrainment Phenomena: Robot Conversation Strategy for Object Recognition. *Appl. Sci.* **2021**, *11*, 2358. https://doi.org/10.3390/app11052358

Academic Editor: Oscar Reinoso Garcia

Received: 7 January 2021
Accepted: 4 March 2021
Published: 7 March 2021

1. Introduction

For social robots in human society, the ability to recognize an object referred to by users is important, as shown in Figure 1. Such indicated object recognition enables social robots to provide information about the indicated object, and pick up and convey the indicated object [1]. Research and development conducted to improve the performance of the indicated object recognition can be divided into two main approaches: *development* and *interactive*. The development 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 visual tracking of a head, hands and head's orientation for recognizing pointing gestures in object references [2]. Schauerte et al. used image processing to integrate speech and pointing gesture recognition [3], 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 [4].

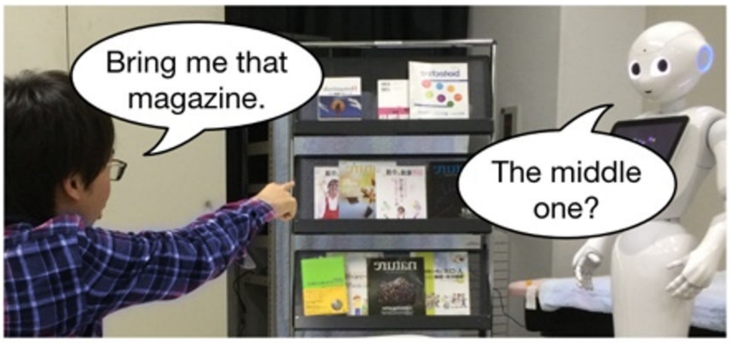

Figure 1. A robot recognizes an object to which a user referred.

Although such developments enhance the sensing capabilities of robots, recognizing the objects indicated by users remains difficult because humans' reference behaviors are often ambiguous during conversations. People enjoy enormous variability in their lexical choices to refer to a thing [5], which degrades the recognition performance because they might not always use the lexical expression contained in a database that stores an object's characteristics, and they do not always use enough words to identify an object [6]. Even if a robot can perfectly recognize speech and/or pointing gestures, they might not be able to identify an indicated object by humans from other objects. How to solve such ambiguity of referencing is an important issue for the indicated object recognition.

The interactive approach aims to improve the performance by decreasing such diversity and the ambiguity of referencing through human–robot interaction. For example, if a robot explains how to refer to an object, people can know a way of referencing and the ambiguity may reduce. The interactive approach tries to change human behaviors to be more favorable for recognition; as a result, the recognition performance could improve. The interactive approach is considered to be useful because it does not require any special equipment and can be used together with the development approach. In this study, we focus on the interactive approach and propose a new strategy of the interactive approach.

1.1. Implicit and Explicit Dyad: Two Types of Interaction Strategies to Clear the Ambiguity of Referencing Behavior

A simple strategy of the interactive approach is the explicit request. Based on the premise that humans refer to an object in an ambiguous way, some researchers have proposed an interactive strategy in which a robot explicitly asks users to provide additional information to identify an indicated object [1,7–9]. 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 for the ambiguous references [1]. Kuno et al. proposed a robotic system that requested users to provide additional information about an object that was difficult to detect [8]. The explicit request strategy focused on how to handle the ambiguity of references and decreased the ambiguity through additional information by users.

In contrast, some previous studies have suggested a strategy that changes ambiguous reference behavior into a clear one without explicit request. One concept leading to the strategy is entrainment. In human–human communication, humans curb such vast lexical choices and elicit terms through the phenomenon of entrainment; when humans are talking, they tend to adjust with an addressee such styles as lexical expressions [10–12], syntax [13], body movement [14–16], etc., and this phenomenon is known as entrainment, also known as alignment. The finding of entrainment inspired researchers of automatic speech recognition (ASR) improve the systems performance [17–20]. Fandrianto and Eskenazi proposed a strategy that changed ASR speaking styles and aligned users' speaking styles, shouting and hyper articulation that decreased ASR performance, with the desired one for the performance [19]. While such studies focused on speech information in the field of ASR, they did not treat gesture information, such as pointing gestures that is useful information to recognize an indicated object. It remains unknown how to incorporate lexical and gestural entrainment to the robot conversation strategy for the indicated object recognition, and whether the strategy improves the recognition performance. In this study, we utilize the stance of using entrainment and propose a novel robot conversation strategy incorporating entrainment phenomenon, we call the proposed strategy implicit entrainment, to improve the performance of indicated object recognition, which implicitly clarifies users' indications through human–robot interaction.

Related studies using explicit request argued that explicit request was useful to decrease the ambiguity of referencing; however, it remains unknown which strategy, implicit entrainment and explicit request, more effectively clarifies the ambiguity of referencing and improves performance. Additionally, social robots should select a more appropriate strategy by considering not only the recognition performance but also the user's impression of the interaction. Even though the performance might be increased by either strategy, such performance improvement becomes non-appropriate if the users feel discomfort during interaction with a robot by the selected strategy and vice versa. In this study, we will compare the two types of strategies, implicit entrainment and explicit request, and discuss the better conversation strategies in the context of indicated object recognition.

1.2. Research Questions

In this study, we take the stance of the interactive approach and propose a novel robot conversation strategy that exploits the entrainment phenomenon and clarifies the human's indication without explicit requests from the robots: the robot uses confirmation behavior to implicitly align with the people's reference behaviors by eliciting clear referencing behaviors from a user. In addition, we compare the proposed strategy with the explicit request strategy, which is the existing strategy of the interactive approach proposed by some past research works. This research answers the following two questions:

1. Does the implicit entrainment strategy that exploits entrainment phenomenon improve the performance of indicated object recognition? (Experiment 1)
2. Robots can employ either a conversational strategy that explicitly requests additional information or one that encourages implicit human entrainment. Which strategy results in better indicated object recognition, and which strategy do humans prefer the robot use? (Experiment 2)

This paper is an extended version of previous works [21–24]. Refs. [21,22] reported the concept and the results of preliminary experiments. In addition, parts of this work were published in [23,24], written in Japanese. To clearly present the novelty and novel knowledge which the paper includes, we have added additional results, references, and more discussions.

2. Background

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

2.1. Lexial Entrainment

When communicating, humans tend to mimic lexical expressions that resemble those of their interlocutor [10,12,13]. This phenomenon is called lexical entrainment. Lexical entrainment is observed not only in human–human interaction but also in human-computer interaction [25,26] and human–robot interaction [27,28]. For example, Brennan reported that humans adopted the computer partner's terms in Wizard-of-Oz experiments using a database query task [26]. Iio et al. verified that humans come to use terms and its categories used by the robot in experiments in which a human referred to an object in conversations with a robot [27]. Brennan and Clark [12] explained lexical entrainment as a consequence of conceptualizations of a "common pact," namely finding common ground [29]. Hence, the entrainment process is affected by the dialogue context and interaction purpose. Garrod and Pickering [30] argued that "dialog is a joint action in which the participants' goal is to establish aligned representations of what they are talking about".

These previous research studies suggest 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 Entrainment

Gestural entrainment has been observed where a speaker tends to adapt their gestures to a partner's gestures in conversations. For instance, Charny reported that the postures of a patient and a therapist were congruent in psychological therapy [31]. In human–robot interaction studies, gestural entrainment is also reported. Breazeal observed that humans aligned their body posture and head pose with those of a robot in conversations [32]. Ono et al. investigated human–robot communications involving guiding route directions and observed that the more a robot uses gestures, the more humans use gestures by entrainment [33]. Iio et al. showed that people used more pointing gestures when a robot used gaze and pointing gestures [34].

The findings of these research studies suggest that, through entrainment, human gestures increased as robot gestures increased.

2.3. Entrainment Inhibition

Some related studies reported cases where entrainment became inhibited in conversations. Shinozawa et al. investigated how humans referred to books when they asked a robot to take them. They changed robot confirmation behaviors that were composed of different amounts of information that were useful to identify the indicated object and investigated how humans refer to a book depending on the confirmation behaviors. They reported that if a robot confirmed the indicated object with precise information, humans' reference speech became diverse, e.g., the reference behaviors with robot confirmation comprised of books' title were more ambiguous than of reference term [6]. Holler and Wilkin reported that mimicking co-speech gestures inhibited lexical entrainment [35]. In their experiment, two interacting participants used a speech and a corresponding co-speech gesture in their first reference to an object; their lexical expression of speech became less precise in their second reference despite consistent co-speech gestures. This finding suggests that gestural entrainment inhibited entrainment of corresponding speech.

These research studies suggest 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 Conversation Strategy Exploiting Entrainment 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 entrainment phenomena, the following paragraphs summarize the approaches and their reasons to elicit a redundant reference from humans:

Lexical Entrainment

- 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 Entrainment

- 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.

Entrainment 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 when the pointing gestures are useful 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 entrainment inhibition. Robots could thus elicit a redundant reference, and the performance of indicated object recognition could improve.

3. Interaction Design

3.1. Object Reference Conversation

For investigating the effect of the proposed strategy exploiting entrainment phenomena, we used an interaction called object reference conversation, that are already being used in human–robot interaction research fields to explore lexical entrainment and humans' reference behaviors [6,27]. Such conversation focuses 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. 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*).

We adapted the proposed conversation strategy to confirmation behavior in the object reference conversations and adjusted the confirmation behavior. The implementation of the confirmation behavior is described in Section 4.3.1.

3.2. Implicit Entrainment and Explicit Request

This section describes the design of implicit entrainment and explicit request in our study. The difference between the implicit entrainment and the explicit request is whether a robot provides specific information that is needed to recognize objects to people to resolve the ambiguity of humans' reference behaviors. As follows, we describe the detailed design of the implicit entrainment and the explicit request.

3.2.1. Implicit Entrainment

We embraced the proposed implicit entrainment strategy described in Section 2.4. This strategy intends that humans come to use references that include enough information to identify the objects by reducing the entrainment inhibitions. The implicit entrainment strategy is implemented in the *Confirm* part in the object reference conversation. Figure 2 shows an example of object reference conversations with an implicit entrainment strategy.

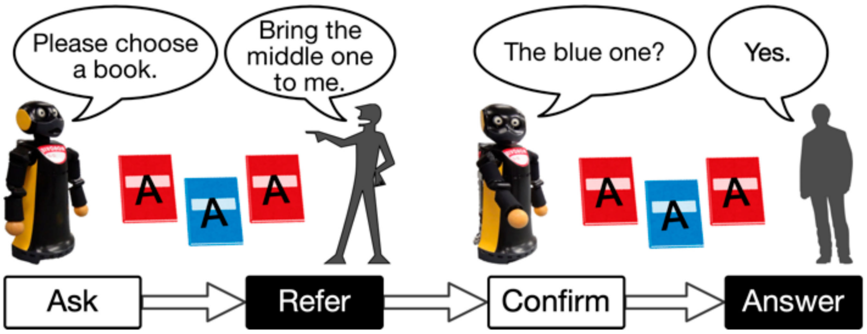

Figure 2. Example of the object reference conversation using the implicit alignment strategy.

3.2.2. Explicit Request

An explicit request is a request that directly asks the specific information that robot wants to the interlocutors. Since it can limit the references, it is expected to reduce unexpected references. If an interlocutor refers to an object in agreement with the robot's requests, the robot will likely recognize it with a high performance. In the explicit request, 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 reference includes sufficient information, the chances of the robot recognizing the referenced object will increase under the poor conditions such as noisy environment, insufficient speech volume, or unclear pointing gestures. In addition, if the interlocutor does not follow the robot's requests, the robot should ask the interlocutor to use the requested information in the object references. This request reminds the interlocutor of the robot's requests and encourages them to use all the information in subsequent conversations. Figure 3 shows an example of object reference conversations using the explicit request strategy.

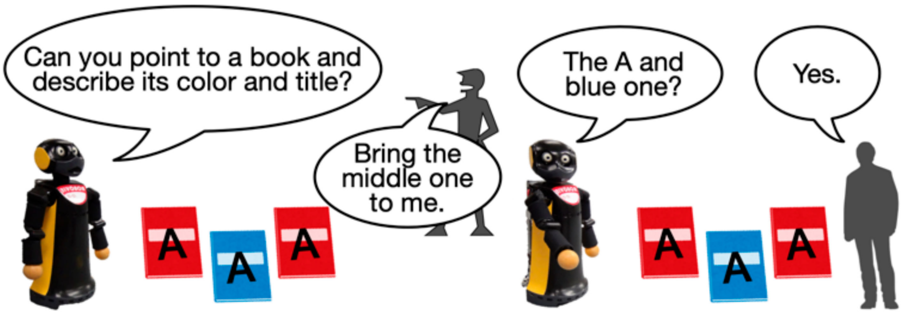
Can you point to a book and describe its color and title?
Bring the middle one to me.
The A and blue one?
Yes.
A
A
A
A
A
A

Figure 3. Example of the object reference conversation using the explicit request strategy.

In summary, we designed the explicit requests in object reference conversaoins as follows: A robot asks the interlocutors to make a reference that includes as much information as possible and requests the interlocutor to use any information that was missing from previous references.

4. System Design

Figure 4 shows the overview of our developed system. The system consists of the following components: sensors, an indicated object recognition function, an object information database, and a conversation strategy selection function. The system will first detect the positions and attributes of the objects arranged in the environment and then saves them in the object information database before an object reference conversation. When a human indicates the object by speech and pointing gestures, the speech recognition module extracts lexical expressions 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. Finally, 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 implicit entrainment or explicit request generation module.

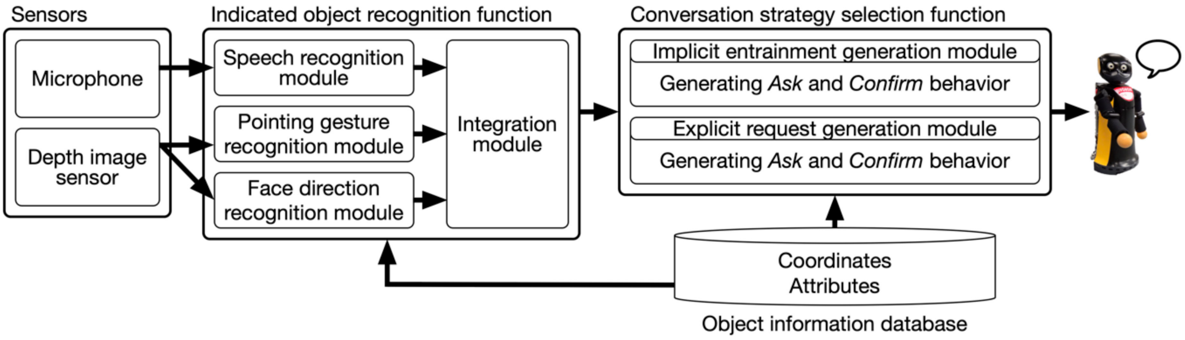
Sensors
Microphone
Depth image sensor
Indicated object recognition function
Speech recognition module
Pointing gesture recognition module
Face direction recognition module
Integration module
Conversation strategy selection function
Implicit entrainment generation module
Generating *Ask* and *Confirm* behavior
Explicit request generation module
Generating *Ask* and *Confirm* behavior
Coordinates
Attributes
Object information database

Figure 4. System architecture to recognize an object to which a user referred and to make the *Ask* and *Confirm* behavior of the robot.

4.1. Robot

We used Robovie-R ver.2, a humanoid robot created by the Intelligent Robotics and Communication Labs, ATR. Robovie-R ver.2 is 1100 mm tall, 560 mm wide, 500 mm deep, and weighs about 57 kg and has a human like upper-body which is designed to communicate with humans. Three DOF are installed 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 [36].

4.2. Indicated Object Recognition

4.2.1. 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 object attributes in the human speech captured by a microphone attached to the human's collar. As for the speech recognition part we used Julius, which has a good recognition performance in Japanese [37].

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, whether the number of n (1, 2, . . . , n) objects are arranged, the likelihood of the object h based on speech recognition sh is calculated as follows:

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

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

4.2.2. Pointing Gesture Recognition Module

The pointing gesture recognition module calculates the normalized reference likelihood of each object based on pointing gesture recognition. This module uses the body frame features captured by a depth image sensor, Kinect for Windows v2, and the object's position data stored in the object information database. The likelihood is modelled 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 (Figure 5) 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.

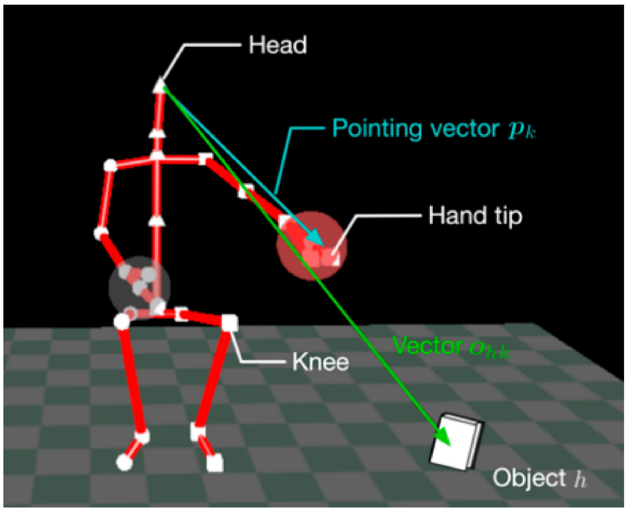

Figure 5. Pointing vector and the vector connecting a human head and an object.

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

$$\alpha_{hk} = \arccos \frac{\boldsymbol{p}_\mathrm{k} \cdot \boldsymbol{o}_\mathrm{k}}{|\boldsymbol{p}_\mathrm{k}||\boldsymbol{o}_\mathrm{k}|} \tag{2}$$

$$g_{hk} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{{\alpha_{hk}}^2}{2}\right) \tag{3}$$

$$p_{hk} = \frac{g_{hk}}{\sum_{h=1}^{n} g_{hk}} \tag{4}$$

Here $\boldsymbol{p}_\mathrm{k}$, and $\boldsymbol{o}_\mathrm{hk}$ indicate the pointing vector and the vector connecting a human head and an object, respectively, on the k-th data with a pointing gesture. g_{hk} indicates the probability that the object h is pointed at, and the angle α_{hk} between $\boldsymbol{p}_\mathrm{k}$ and $\boldsymbol{o}_\mathrm{hk}$ 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 (5):

$$p_h = \begin{cases} 0 & (m\ =\ 0) \\ \frac{\sum_{k=1}^{m} p_{hk}}{m} & (m\ >\ 0) \end{cases} \tag{5}$$

4.2.3. 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 on a level surface and the vector connecting a head and an object (Figure 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 [38]. The likelihoods are normalized from 0 to 1.

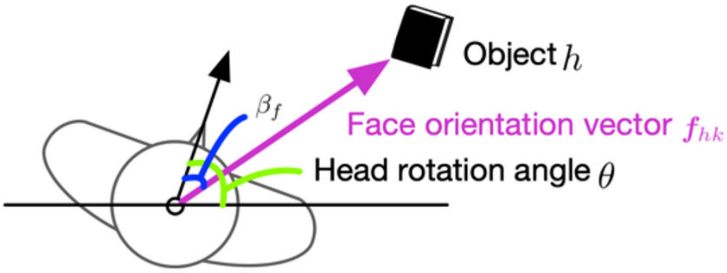
Object h
β_f
Face orientation vector f_{hk}
Head rotation angle θ

Figure 6. Vector and angles used to identify face direction.

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:

$$f_{hk} = \frac{w_{hk}}{\sum_{h=1}^{n} w_{hk}} \tag{6}$$

$$w_{hk} = \begin{cases} 0 \ (\beta_{hk} \geq 11\pi/18) \\ 1 \ (\beta_{hk} < 11\pi/18) \end{cases} \tag{7}$$

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 (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:

$$f_h = \begin{cases} 0 & (j = 0) \\ \frac{\sum_{k=1}^{j} f_{hk}}{j} & (j > 0) \end{cases} \tag{8}$$

4.2.4. 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:

$$l_h = \frac{s_h + p_h + f_h}{\sum_{h=1}^{n}(s_h + p_h + f_h)} \tag{9}$$

The integration module obtains the object $o_{\max}$ with the highest likelihood of objects in the environment using Equation (10) and estimates the object as indicated object by a person:

$$o_{\max} = \operatorname{argmax}_{x \in \mathrm{h}}(l_x) \tag{10}$$

4.3. 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 implicit entrainment strategy or the explicit request strategy is used. The implementation of each strategy is described below in detail.

4.3.1. Implicit Entrainment Generation Module

When the robot uses the implicit entrainment strategy, the robot does not explicitly request how to refer to an object in the *Ask* part. The implicit entrainment 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 entrainment strategy, and the robot confirms the indicated object with minimum information for distinguishing among objects based on the proposed design described in Section 2.4. The procedure comprises two steps: (1) Deciding whether to use a pointing gesture, and (2) deciding which object attributes to use in the confirmation speech.

Deciding whether to Use a Pointing Gesture

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 around the indicated object adjacently and a pointing gesture does not narrow the candidates for the indicated object, 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 deciding 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. [39] 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 [39]. 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 [40,41], is a maximum of $\pi/9$ rad [42]. 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.

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 7 shows an example of how to decide whether to use a pointing gesture. In Figure 7a, 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 7b, 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.

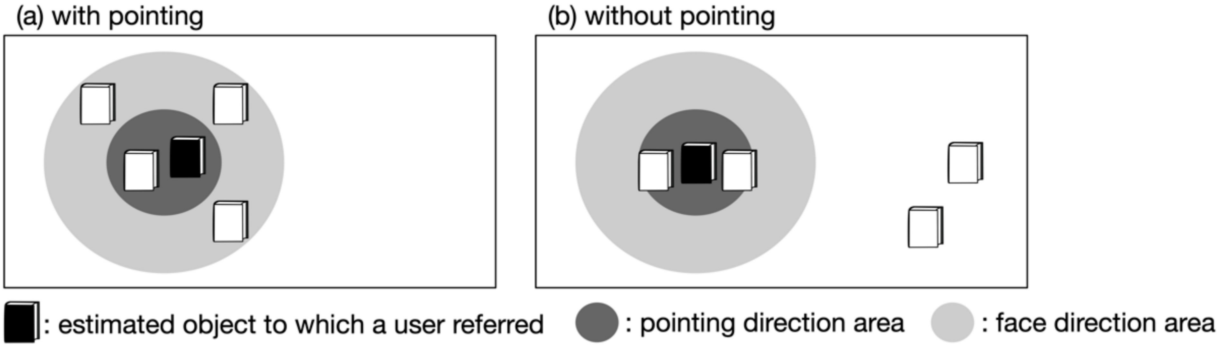
(a) with pointing
(b) without pointing
: estimated object to which a user referred
: pointing direction area
: face direction area

Figure 7. How to decide whether to use a pointing gesture.

Deciding Which Object Attributes to Use in the Confirmation Speech

Next, we describe how to decide the attribute set used in the speech. The robot uses the minimal number of object attributes to identify the indicated object in a confirmation. First, if only indicated object is situated within the pointing or facing direction area, the robot gives one attribute that is chosen randomly. 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. In this case, one attribute is sufficient for identifying the object because a pointing gesture or face direction can distinguish it from the others. If there are other objects within the area, the robot uses enough minimal attributes to identify the indicated object. If there are several sets of minimal attributes, the system needs to select one set. The system selects the set based on the similarity among the attributes of each set and those of objects in the area. The system calculates the similarity and chooses the set with the least similarity. This is because the robot wants to elicit easily recognizable speech from an interlocutor as much as possible. If an attribute set of the reference speech is similar to that of objects in the area, one recognition failure of the attribute would cause a misrecognition of the indicated object. As the 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 string $\mathrm{s}_{\mathrm{one}}$ and s-th character of string $\mathrm{s}_{\mathrm{two}}$ is recursively given by $LD(p,s)$ as follows:

$$LD(p,s) = \min\begin{cases} LD(p-1,\ s)+1 \\ LD(p,s-1)+1 \\ LD(p-1,\ s-1)+c \end{cases} \tag{11}$$

$$c = \begin{cases} 0(char(\mathrm{s}_{\mathrm{one}},p) = char(\mathrm{s}_{\mathrm{two}},s)) \\ 2(char(\mathrm{s}_{\mathrm{one}},p) \neq char(\mathrm{s}_{\mathrm{two}},s)) \end{cases} \tag{12}$$

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.

For example, Figure 8a 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 8b 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”.

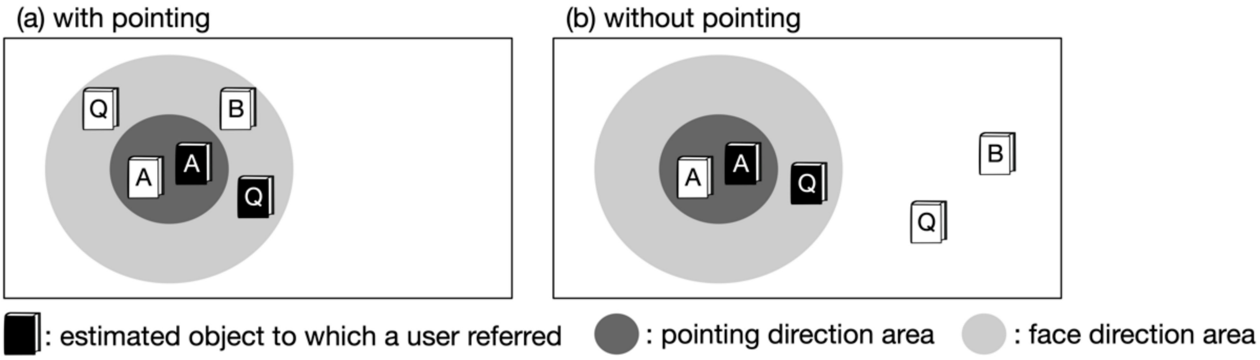
(a) with pointing
(b) without pointing
Q
B
A
A
Q
A
A
Q
B
Q
: estimated object to which a user referred
: pointing direction area
: face direction area

Figure 8. How to decide which object attributes to use in the confirmation speech.

4.3.2. Explicit Request Generation Module

When the robot uses the explicit request strategy, this module decides the *Ask* behavior adopted the explicit request strategy described in Section 3.2.2, and the robot explicitly provides requests about how to refer to objects. The robot explains what kinds of information is needed for the robot recognition by showing the reference way including all object attributes to an interlocutor. 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 did not 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 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 color.”

Figure 9 shows the procedure to generate the explicit requests. First, the explicit request generation module judges whether the object reference conversation is first time. If the conversation is first time, an explicit request is the speech that requests a reference that includes object attributes that the robot can recognize, pointing gestures and face direction. If the conversation is second time or later, the module judges whether requested attributes and a pointing gesture were included in the previous reference of an interlocutor based on the speech and pointing gesture recognition results.

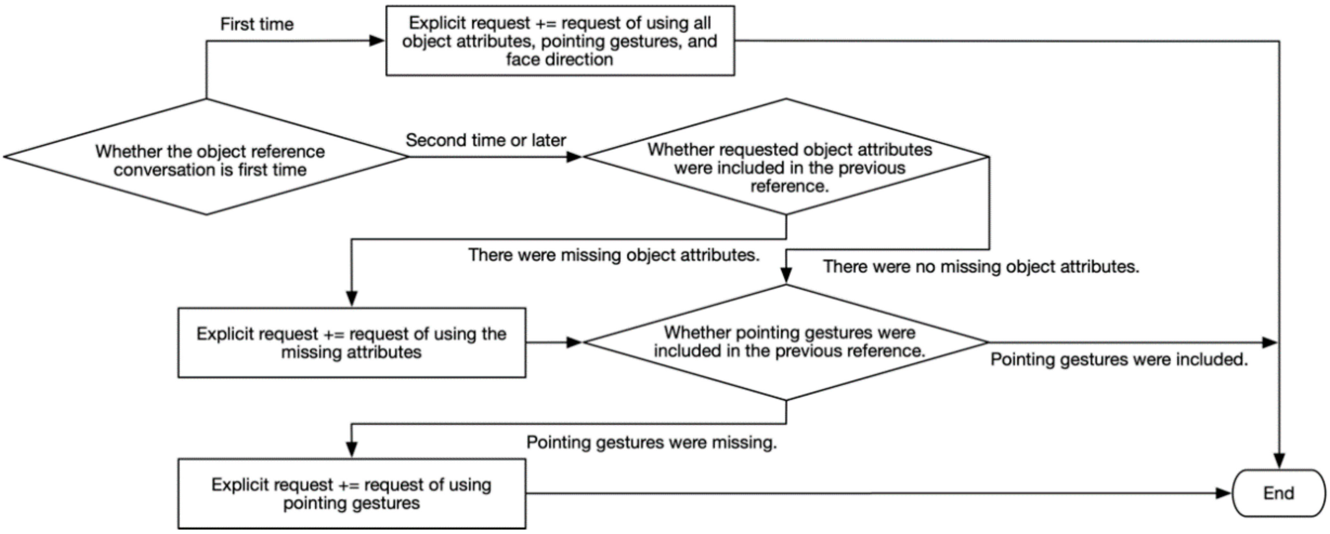
First time
Explicit request += request of using all object attributes, pointing gestures, and face direction
Whether the object reference conversation is first time
Second time or later
Whether requested object attributes were included in the previous reference.
There were missing object attributes.
There were no missing object attributes.
Explicit request += request of using the missing attributes
Whether pointing gestures were included in the previous reference.
Pointing gestures were included.
Pointing gestures were missing.
Explicit request += request of using pointing gestures
End

Figure 9. Procedure to make an explicit request.

If an object $\mathrm{o_p}$ was indicated by an interlocutor in the previous conversation and reference likelihood based on a pointing gesture recognition is zero ($p_{\mathrm{o_p}} = 0$), this means that an interlocutor referred to the object without a pointing gesture, the 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 objects included in the interlocutor's previous reference based on the speech recognition results. If there is a missing attribute, 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 robots confirm an object in the same way the robot requested to an interlocutor. Specifically, the robot confirms an indicated object o_{max} by saying all the object's attributes $O_{o_{\mathrm{max}}}$ with a pointing gesture.

5. Experiment 1

Through Experiment 1, we investigate whether the proposed implicit entrainment strategy elicits a redundant reference and improves the performance of indicated object recognition. We compared performance of a system adopting the implicit entrainment strategy (proposed system) with that of a system non-adapting the implicit entrainment strategy (alternative system). We basically employed the developed system described in Section 4 excluding explicit request generation module for both proposed and alternative system; however, in the *Confirm* part, the alternative system did not adopt proposed strategy and confirm an indicated object in a way that did not consider the entrainment phenomena.

5.1. Experiment Settings

5.1.1. 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 entrainment, 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 entrainment 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 1-1: 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.

5.1.2. Conditions

To verify the hypotheses, we controlled the robot's confirmation behavior and the arrangement of objects in the environment.

The confirmation factor was a within-participants design and had two levels: minimum attributes and all attributes. In the minimum attributes level, the robot confirms the indicated object with minimum information for identifying the indicated object among objects. In the all information level, the robot confirmed the objects with all the information, it gave every attribute of an object and pointed during the confirmations. The speech format of the confirmations was the sequence of the object attributes. For example, the robot said, "That circle and red book?" or "That triangle, B and blue book?" In this experimental design, in the confirmation factor, we did not separate the speech from the pointing gesture control as a level. This is because the content of robot's speech is calculated depending on the result of pointing gesture control, and strict separation of speech and pointing gesture control is difficult. In addition, even if the speech level is fixed to the minimum information level, it is difficult to distinguish the effects of a robot's gesture itself from the effects of combination of a 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.

The arrangement factor was a within-participants design and had three levels: sparse set, two groups and congestion. We set this factor considering the influence of the arrangement of objects on the performance of the indicated object recognition because the arrangement may affect what kinds of lexical expressions and pointing gestures are chosen by people. For example, if objects were sparsely arranged, humans would refer to an object by pointing gestures and a deictic. On the other hand, if objects were densely arranged, humans would refer to an object by a speech that is composed of many object's attributes because only use of pointing gestures is not enough to identify the object. Hence, not only the robot's confirmation behavior but the arrangement of objects would have influence on humans' reference behavior, and the performance of the indicated object recognition would change depending on the arrangement.

In the experiment, we asked participants to select objects and arrange them freely under these three conditions. For example, we instructed them to "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 10 shows the examples of objects arrangement by a participant.

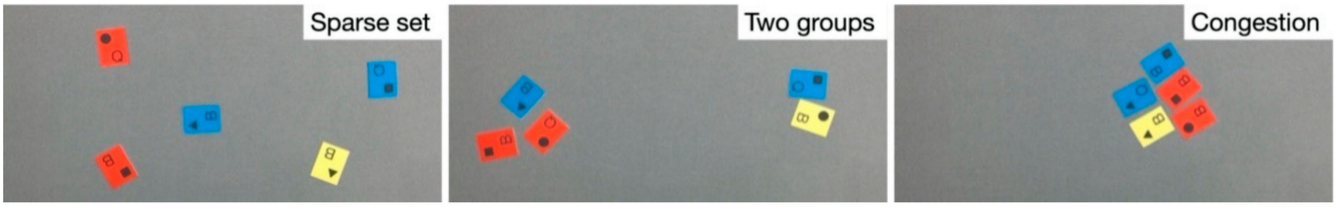

Figure 10. Example of book arrangements by a participant.

5.1.3. Environment

Figure 11a shows the experimental environment. Our participants sat in front of the robot. The position of the robot and the char were fixed to keep the distance between the robot and the participant, and to eliminate the distance effects on the participant's object reference behavior. Objects were arranged without overlapping in a 1.5 m by 3.3 m rectangular area between the robot and the participant. We placed camcorders behind the robot and on the right side of the robot respectively to record the experiment (Figure 11b).

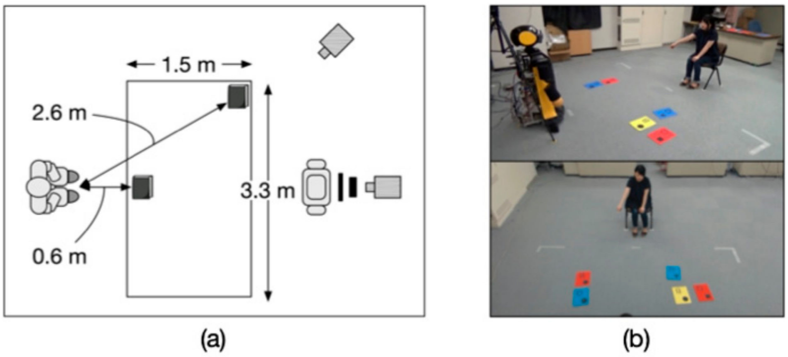

Figure 11. (**a**) Experimental environment; (**b**) example scenes in Experiment 1.

As the objects that participants refer to, we prepared a book that was 21 cm by 27.5 cm, and their attributes were color, symbol, and a letter on the cover. There were three colors: red, blue, or yellow. Three symbols were placed on the book covers: a circle, a triangle, or a square. There were two letters: Q and B. We prepared 18 books to satisfy all combinations of attributes. The attributes and positions of arranged books were automatically recognized by using an image from calling mounted RGB camera. The recognized attributes and positions were stored in the object information database.

5.1.4. Procedure

The procedure of Experiment 1 was as follows: First, the participants were given a brief description of the experiment procedure. Next, we gave them the following instructions: "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 dealing with a person." We did not give them the instruction about how many times they referred to each object, the order of reference to objects and what kinds of expression can be used for a reference.

After these instructions, the participant chose five books among the 18 and placed them based on the arrangement levels. The participants repeated the object reference conversation 10 times because to verify the influence of robot's confirmation behavior, it is necessary to have the conversation multiple times. We decided the number of the conversation based on the related work that verified human–robot entrainment [27,43]. The example of the conversation was as follows. First, the robot said, "Please choose a book," and the participants freely referred to an object in the environment. The robot estimated the indicated object by indicated object recognition function and confirmed the object by the way of confirmation calculated by confirmation selection 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 same as the indicated object, the participant answered "No, it isn't," to the robot's confirmation. In both case of the answers, the robot did not reply to the participant's answer and the conversation continued.

We call 10 object reference conversations sessions, which were conducted in every arrangement condition: sparse set, two groups, and congestion. They eventually conducted two by three sessions with different confirmation levels. Participants re-selected five books and re-arranged 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.

5.1.5. Measurement

We measured the success rate of indicated object recognition per session. We calculated it from the success case that the book confirmed by the robot was same as the indicated book by a participant: a participant answered "Yes, it is," to the robot's confirmation. However, we should consider the case that although the robot correctly confirmed the indicated object, the participant answered "No, it isn't," and although the robot mistakenly confirmed the non-indicated object, the participant answered "Yes, it is." Hence, the experimenter checked the existence of such error cases by the videos from camcorders and the ceiling-mounted RGB camera (Figure 11b) and recorded speech sound. As a result, two of 1440 object reference conversations were such error cases. First, although the robot confirmed the indicated object, the participant answered "No, it isn't." Second, although the robot confirmed the non-indicated object, the participant answered "Yes, it is." We calculated these conversations as a success and an error, respectively.

5.1.6. Participants

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

5.2. Results

5.2.1. Verification of Prediction

Figure 12 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 = 0.037, η_p^2 = 0.176). This result showed success rate of indicated object recognition with minimum information level was significantly higher than the that with all information level. Thus, Prediction 1-1 is supported. We found no significant main effects in the arrangement factor (F(2,46) = 2.245, p = 0.117, η_p^2 = 0.089). We also found no significant interaction between the two factors (F(2,46) = 2.659, p = 0.081, η_p^2 = 0.104).

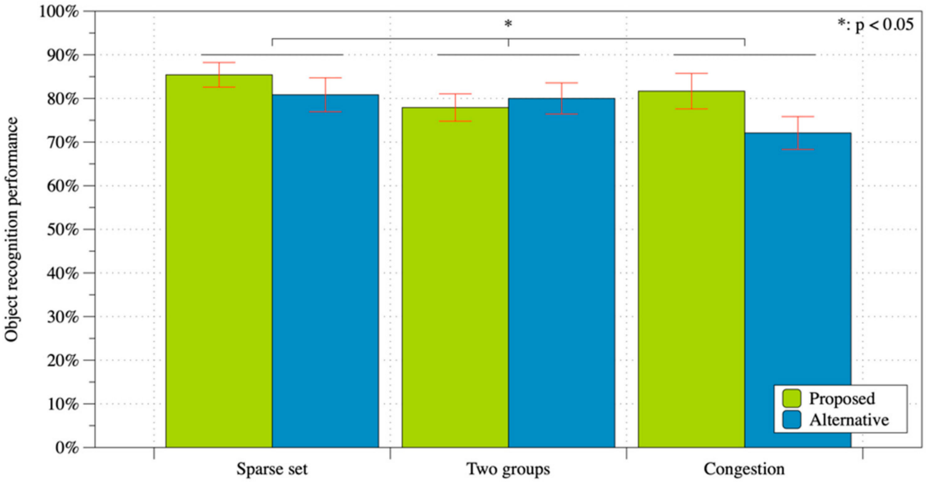

Figure 12. Performance of object reference recognition with standard error of Experiment 1.

5.2.2. Number of Object Attributes and Pointing Gestures in Humans' Reference Behaviors

To investigate whether the number of object attributes and pointing gestures in references changed depending on the robot's confirmation behavior, we measured the mean number of object attributes in the speech and pointing gestures per sessions. Tables 1 and 2 show the mean number of object attributes and the mean number of pointing gestures respectively. We defined the number of pointing gestures from with or without pointing gestures. If a participant referred to an object with pointing gestures, the number of pointing gestures was one. Similarly, if a participant referred to an object without pointing gestures, the number of pointing gestures was zero. We calculated the mean number of pointing gestures per one session that means 10 object reference conversations. If a participant referred to an object with pointing gestures in 10 object reference conversation of one session, the mean number of pointing gestures was one.

Table 1. Mean number of object attributes with standard error.

	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)

Table 2. Mean number of pointing gestures with standard error.

	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)

First, we conducted a two-factor repeated measure ANOVA for the mean number of object attributes. Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(2)$ = 16.8, p = 0.011), therefore, degrees of freedom were corrected using Greenhouse–Geisser estimates of sphericity (ε = 0.747). We found significant main effects in the arrangement factor (F(1.495,34.384) = 5.026, p = 0.011, η_p^2 = 0.179). Post hoc comparisons using the t-test with Bonferroni correction indicated that significant difference was revealed between the sparse set level and the congestion level (p = 0.048), and between the two groups level and the congestion level (p = 0.035). In other words, in the environment where objects were arranged close to each other, participants tended to refer to an object with speech containing more attributes of an object than in the other environment of books arrangement. Main effects in the confirmation factor was not revealed (F(1,23) = 1.120, p = 0.301, η_p^2 = 0.046). Interaction between confirmation and arrangement factor was also not revealed (F(2,46) = 1.000, p = 0.376, η_p^2 = 0.042).

Second, we conducted a two-factor repeated measure ANOVA for the mean number of pointing gestures. Main effects were not revealed in the confirmation factor (F(1,23) = 0.956, p = 0.338, η_p^2 = 0.040) and arrangement factor (F(2,46) = 2.775, p = 0.073, η_p^2 = 0.108). In other words, significant influence on the number of pointing gestures by the robot's confirmation behavior and the arrangement of objects was not observed.

5.2.3. Change of Referencing Style through Interaction

We proposed the robot's confirmation behavior to entrain the redundant reference based on the consideration the humans' desirable reference behavior to improve the performance of indicated object recognition is the redundant reference, that is the reference behavior with the speech that contains as much information as possible that is valuable to identify the object and with pointing gestures as described in Section 2.4. To verify whether the frequency of the such redundant reference depending on the robot's confirmation behavior, we analyzed the humans reference behavior from the point of view of reference redundancy of speech and with or without pointing gestures.

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 required to uniquely identify the indicated objects in the environment. The objects we used in the experiment have three attributes: color, symbol, and letter. Hence, the number of attributes for the object in the participant's references was defined as 0–3. For example, if the participant did not referred to any of the attributes, the number of object attributes is 0, while the participant referred to all of the attributes (i.e., color, symbol, and letter), the number of object attributes is 3. In the environment, the minimum number of attributes required to uniquely identify the indicated object ranges from 1 to 3. Accordingly, reference redundancy ranges from -3 to 2. Suppose a participant refers to a book with no attributes (e.g., "that book" or "this book") in the environment where all of the attributes are needed to uniquely identify objects. In this case, minimum number of attributes is 3 and, thus, the reference redundancy is -3. Likewise, if a participant refers to a book with all three attributes in the environment where there is one red book and only color is needed to uniquely identify the object, the minimum number of attributes is 1, and the reference redundancy is 2.

The redundant reference is defined as the reference that of reference redundancy of speech is more than 0 and with a pointing gesture, and we measured the rate of the redundant reference per sessions, that is 10 object reference conversations. Figure 13 shows the results. We conducted a two-factor repeated measure ANOVA and main effects of the confirmation factor were close to significance (F(1,23) = 3.641, p = 0.069, η_p^2 = 0.137). No effect was observed in main effects of the arrangement factor (F(2,46) = 0.727, p = 0.489, η_p^2 = 0.031) and interaction between confirmation × arrangement factor (F(2,46) = 0.312, p = 0.734, η_p^2 = 0.013). These results suggest the trend 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.

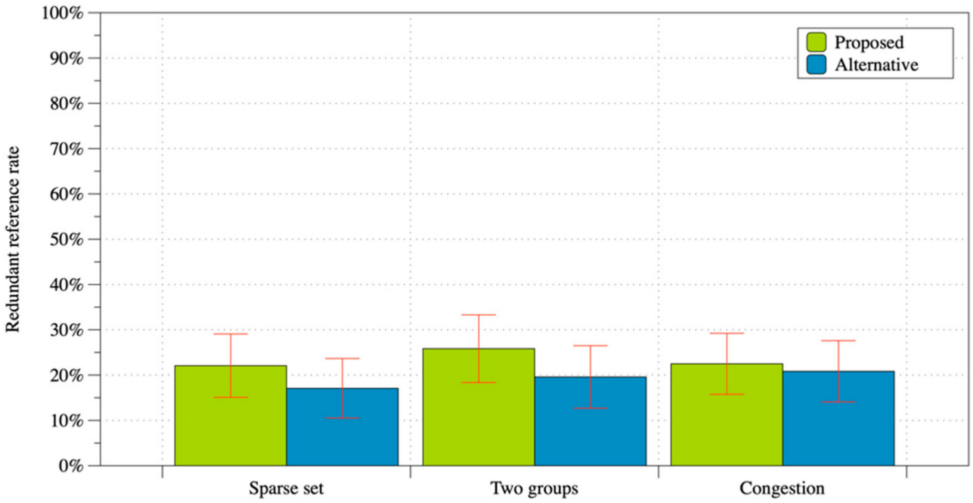

Figure 13. Rates of redundant reference per session.

5.3. Discussion

5.3.1. Implication

In Experiment 1, we verified whether the performance of indicated object recognition improves through the robot's confirmation behavior. The experimental results show that the performance of indicated object recognition significantly improves by the robot's confirmation behavior with minimum information to distinguish the indicated object in the environment. The main contribution of this study is that we verified the performance of indicated object recognition could change only through the change of robot's confirmation behavior. Our contribution is useful to design human–robot interaction. This is because confirmation behavior is often observed in human–human interaction and the confirmation is easy to apply to the robot's behavior model to interact with people. In addition, the proposed method changing the confirmation behavior, does not depend on a speech recognition method, algorithm of recognizing reference behavior, a variety of sensors, and so on, and could be easily applied to existing robotic system for indicated object recognition.

5.3.2. Influence of System Parameters on the Results

We decided the tolerance for the pointing gesture and the face direction based on the related works. However, such tolerance would depend on the appearance of the robots and/or humans. For example, in the case where a robot's body is larger/smaller and a robot can change the line of sight, we would need to adjust the tolerance. Similarly, the parameters to decide whether to use a pointing gesture might depend on the size of objects and the largeness of an environment. In each recognition modules, we also decided the parameters such as a humans' field of view based related works. Such parameters have a certain level of generality if a robot's conversation partner is a human. However, such parameters might also need to adjust. For example, because of cultural differences, we need to change the calculation method of the speech similarity in generating confirmation speech, we used Levenshtein distance in this study.

6. Experiment 2

The results of Experiment 1 showed that the proposed implicit entrainment strategy elicited a redundant reference from users and the recognition performance improved. Through Experiment 2, we compared the proposed implicit entrainment strategy with the explicit request strategy which is proposed by some previous studies, and we investigate whether the implicit entrainment or explicit request strategy is better for object recognition contexts in conversations with people. In this experiment, we employed the developed system and experimentally compared the two strategies by switching adapted strategy in the conversation strategy selection function.

6.1. Experiment Settings

6.1.1. Hypothesis and Prediction

The interlocutor might not be certain about how to refer to the object if there is no explicit request concerning the way of the references, while if the robot explicitly requests a particular way of reference, the interlocutor will know how to refer to an object and may use it in conversation. The references following the robot's explicit request leads to more accurate recognition of the indicated object. However, referencing an object in an explicit manner in daily conversations is redundant and, thus, the interlocutor might think the conversation unnatural. Similarly, the interlocutor might feel frustrated, since the conversation is troublesome. Based on these considerations, we make the following two predictions:

Prediction 2-1: The indicated object recognition performance of conversations using the explicit request strategy will be better than that of conversations using the implicit entrainment strategy.

Prediction 2-2: Conversations using the implicit entrainment strategy will be perceived as more positive by the interlocutor than conversations using the explicit request strategy.

6.1.2. 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 entrainment. The applied strategy factor was a within-participant condition. In this experiment, we compared the applied strategies for the robot's *Ask* and *Confirm* behaviors and there was no difference in the procedure to recognize the reference behaviors and estimate the indicated object. The design and implementation of the implicit entrainment and explicit request strategy are described in Sections 3.2 and 4.3. We summarize the strategies below.

In the explicit request condition, the robot asks interlocutors to make a reference that includes as much information as possible while the robot encourage the interlocutor to use the information missing in the *Ask* part. In the *Confirm* part, 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 entrainment condition, unlike the explicit request condition, the robot does not explicitly provide requests about the reference type; the robot merely says, "Please choose a book" in the *Ask* part. However, in the *Confirm* part, the robot confirms the indicated object with implicit entrainment strategy. In this condition, the robot confirms the object with minimum information to identify the object among the objects in the environment.

6.1.3. Environment

We used the same environment as shown in Figure 11, described in Section 5.1.3.

6.1.4. Procedure

The procedure of the Experiment 2 was as follows. First, we explained the experiment to the participants and asked them to sign consent forms. Next, we gave them the following 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 implicit entrainment and explicit request, 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 Experiment 1. 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. After placing the books in one place, the participants repeated the object reference conversations 10 times in both applied strategies (implicit entrainment and explicit request). The participants answered questions about their impressions of the conversations after each conversation session.

6.1.5. Measurement

Recognition Performance

The recognition performance represents the success rate of the object reference recognition calculated from the number of object references correctly recognized by the robot in each conversation session which consists of 10 sets of object reference recognitions.

Impression of Conversations

To investigate the participant's impressions of the conversations, we prepared the following eight 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 load (load).
2. The conversation with the robot was troublesome (troublesome).
3. It was easy to make a reference to a book (easiness).
4. The conversation with the robot was difficult (difficulty).
5. The conversation with the robot was natural (natural).
6. The conversation with the robot was easy to understand (understandability).
7. I felt familiarity with the robot (familiarity).
8. Overall impression of conversation (overall impression).

6.1.6. Participants

Twenty-six native Japanese speakers (13 females and 13 males, who averaged 36.5 years of age, *SD* = 9.3) participated in our experiment.

6.2. Results

6.2.1. Verification of Prediction 2-1

Figure 14 shows the recognition performance results. In order to verify the effect of each condition, we conducted a paired t-test and found no significant difference between the two applied strategy conditions (t(25) = -1.06, $p = 0.302$, $d = 0.274$). Accordingly, the result indicated that Prediction 2-1 was not supported.

6.2.2. Verification of Prediction 2-2

Figure 15 shows the results of the questionnaire items. In order 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 (t(25) = 2.440, $p = 0.022$, $d = 0.58$), for the *troublesome* feeling (t(25) = 4.556, $p < 0.001$, $d = 0.89$), for the *natural* feeling (t(25) = -2.403, $p = 0.024$, $d = 0.51$), and for the *overall impression* among the conditions (t(25) = -2.339, $p = 0.028$, $d = 0.36$). No significant difference for the *easiness* (t(25) = -0.814, $p = 0.423$, $d = -0.160$), *difficulty* (t(25) = 1.641, $p = 0.113$, $d = 0.322$), *understandability* (t(25) = -1.383, $p = 0.179$, $d = -0.271$), and *familiarity* (t(25) = 0.137, $p = 0.892$, $d = 0.027$). Overall, the results supported Prediction 2-2.

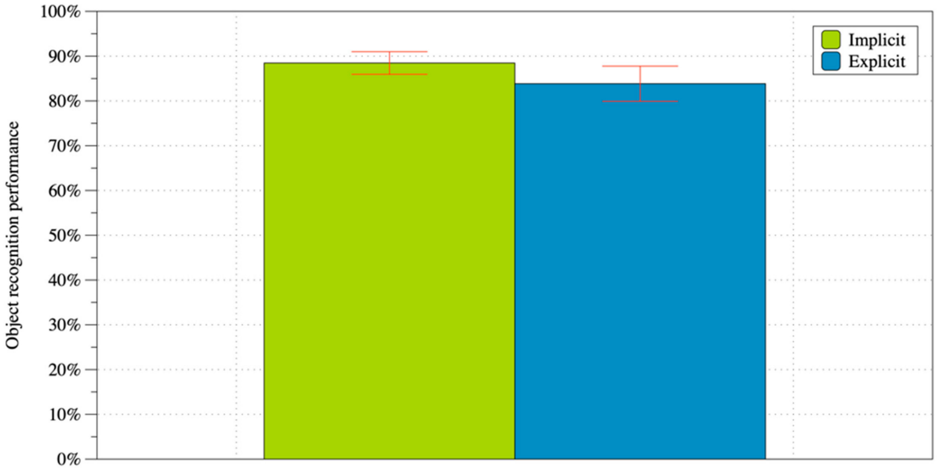

Figure 14. Performance of object reference recognition with standard error of Experiment 2.

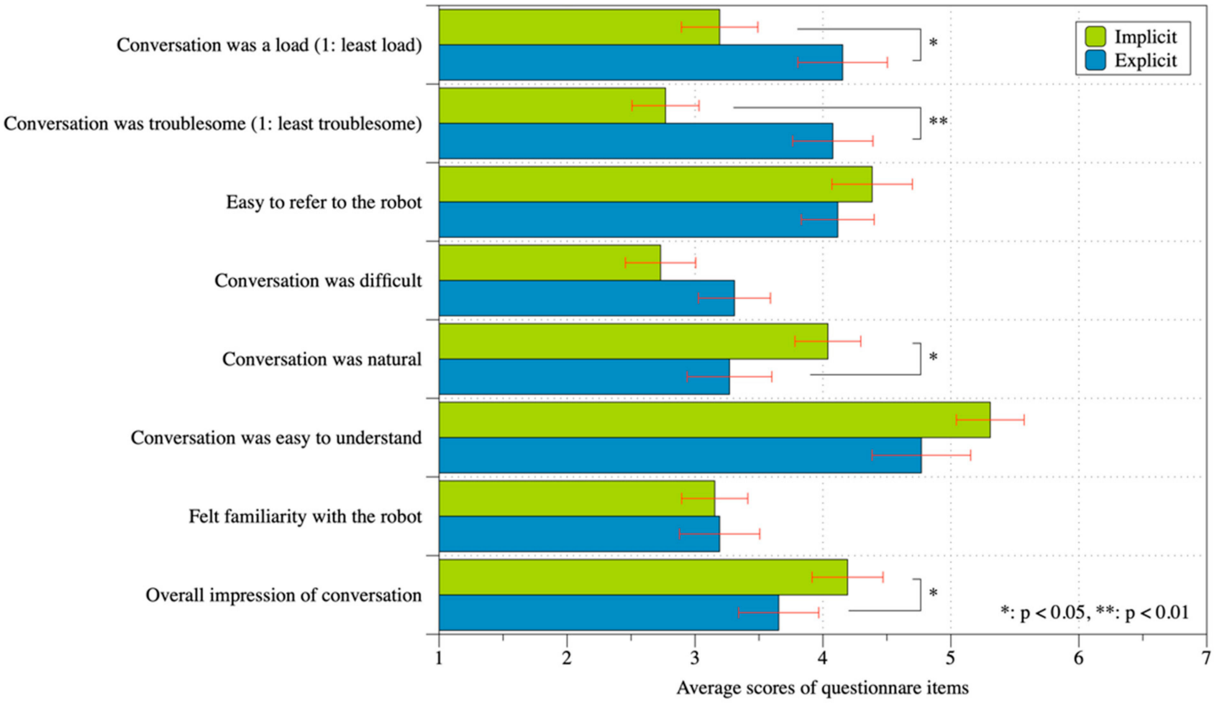

Figure 15. Questionnaire results of conversation impressions with standard error.

6.3. Discussion

6.3.1. Comparison between Explicit Request and Implicit Entrainment

The results indicated that there was no significant difference in the recognition performances between the two strategies. This result suggests two way of interpretations. First, we can suppose that the implicit entrainment strategy enhanced the recognition performance at the same level as the explicit request strategy. Although we directly compared the performance of the implicit entrainment strategy with the explicit request strategy and no baseline for the performance, the implicit entrainment strategy likely enhances the recognition performance, as argued by Experiment 1. Second, the explicit request strategy might not enhance the recognition performance well. While we expected that the interlocutors would refer to an object as the robot requested, in the experiment, 28% of references in the explicit request condition did not follow the robot's requests. Some of participants referred to an object with information that was dropped, such as a pointing gesture and an object's attribute. The explicit request is not likely assured for inducing interlocutors to use clear references.

The experiment results show that the interlocutor's impressions of the conversations with the implicit entrainment 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 entrainment 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. Overall, it is suggested that the implicit entrainment strategy is better than the explicit request strategy for the context of indicated object recognition. We assume that this difference is not subtle if we imagine the repetitive and long-term interaction. In such interaction, one-shot subtle negative impressions are accumulated, and it would make non-negligible differences of impressions. We believe that our findings are useful for designing interactions for social robots; better impressions of conversations are important as well as recognition performance for them because poor impressions cause people to hesitate in interaction with robots.

6.3.2. Comparison of Referencing Style

To compare referencing style between implicit entrainment and explicit request, we investigate (1) the mean number of object attributes and pointing gestures in the references per conversation session and (2) the rate of the redundant reference in the same way as described in Sections 5.2.2 and 5.2.3, respectively.

First, we investigate whether the number of object attributes and pointing gestures in the references changed depending on the robot's conversation strategy. Table 3 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, $p = 0.106$, $d = 0.29$), but a significant difference was found for the mean number of pointing gestures (t(25) = 4.477, $p < 0.001$, $d = 0.19$), with the explicit request eliciting more pointing gestures than the implicit entrainment. These results indicate that humans refer to an object by using a similar amount of object information in speech with both the explicit request and the implicit entrainment, and they refer to an object with more pointing gestures in the conversations with the explicit request than with the implicit entrainment.

Table 3. Comparison of mean number of object attributes and pointing gestures included in references with standard error between explicit request and implicit entrainment strategy.

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

Second, we investigate the rate of the redundant reference per conversation session (ten times object reference conversations). Figure 16 shows the results. The results of a paired *t*-test showed that a significant difference existed in the rate of the redundant reference (t(25) = 5.21, $p < 0.001$, $d = 1.00$). These results indicate that humans use more redundant reference to the robot with explicit request than with implicit entrainment. It is assumed that this significant difference is caused by the significant difference of the number of pointing gestures, because we found no significant difference for the number of object attributes in references.

The rate of redundant reference is significantly different between the two strategies; however, the recognition performance is not significantly different between the two strategies (shown in Figure 12). These results seem that both the explicit request and the implicit entrainment strategies elicit sufficient speech and pointing gestures to identify an object with the same level of accuracy at least in our experimental settings. Even in the conversations with the implicit entrainment 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.

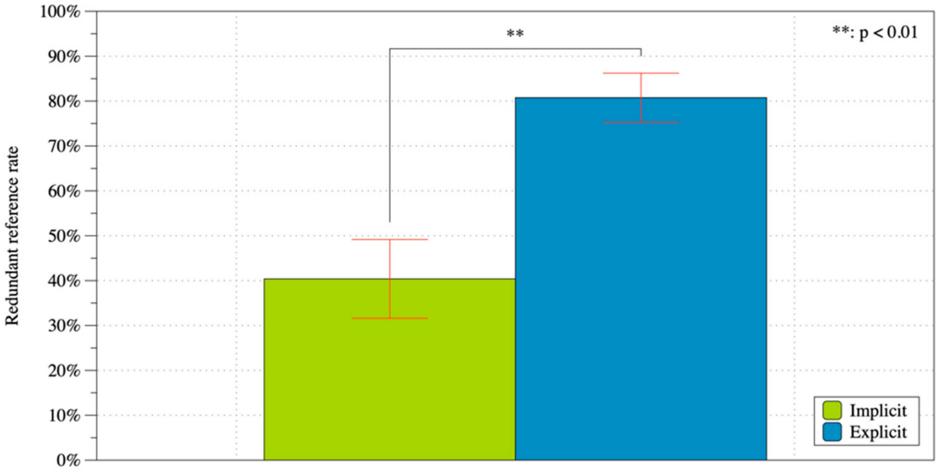

Figure 16. Comparison of rates of redundant references with standard error between explicit request and implicit entrainment.

In Experiment 2, we only used the congestion object arrangement because the recognition performance in the congestion settings is lower than the recognition performance in the sparse set and two groups when the robot did not use the proposed strategy in Experiment 1, and we assumed the congestion arrangement is suitable for comparison between two interactive strategies. However, the comparison between other object arrangements would provide additional findings in terms of the relationships between the interactive conversation strategies and the object arrangements, and the relationships should be further studied in future work.

7. General Discussion

In Experiment 1, we have proposed the strategy incorporated entrainment phenomena and verified that the strategy improved the performance of indicated object recognition. In Experiment 2, we compared the proposed strategy with another interactive strategy that is introduced by some past studies [1,7–9] from a perspective of both recognition performance and conversation impressions. Overall, our experimental results provide a suggestion for interaction design for social robots that need to recognize an indicated object referred by a user: robot behavior can implicitly lead humans' behavior to desirable one for recognition with good impressions compared to a robot behavior that explicitly requests desirable behavior for the robot. We believe that our findings help to design the interaction of social robots because (1) the proposed conversation strategy aligning humans' behavior through interaction is compatible with other recognition techniques such as development of new devices and/or algorithms tracking humans' behaviors, and (2) implicit and explicit dyad is conceivable in other contexts, picking up objects, item delivery, spoken dialogue systems, etc.

On the other hand, we conducted our experiment's study in a limited situation and there are some limitations. First, the participants referred to objects that have a limited number of attributes. In the real environment, the attributes associated with objects are non-limiting and the variety will affect the methods of reference. However, the interaction strategy to clear reference behaviors does not depend on features, and our findings are extendable for other objects. In addition, we used the objects have unique attribute combination between them. In the real environment, there is a situation where objects would have similar attributes and a speech could not distinguish an object with others even if the speech contains all attributes of the object. In such situation, uniquely to identify an object, robots would need to speech with an expression of a position relationship, for example "next to" or "near" to reduce candidates of a confirmation. Additionally, an expression with a place in an environment, for example, "at the corner of the room" or "on the table" would be useful to uniquely identify an object. Humans also use referring expressions that include spatial propositions. Doğan et al. proposed the deep learning method that generates unambiguous spatial referring expressions [44]. To decrease the ambiguity of referring expressions, considering such spatial expressions would also be useful.

In Experiment 1, we did not separate the speech control from the pointing gesture control to treat the entrainment inhibition which is caused by the relationships between lexical and gestural behaviors. Due to this setting, it is difficult to distinguish the effects of lexical entrainment from that of gestural entrainment. To separately understand the effects of lexical and gestural entrainment, we need to control the number of attributes in speech and with/without pointing gestures as experimental conditions, although, in this case, the robot cannot treat the alignment inhibition. It would also be a future work to use non-humanoid type agents, e.g., a robot arm or a smart speaker, to investigate the effects of conversation strategy utilizes entrainment phenomena. Since, if a robot has different capabilities of sensors/actuators, the expected capability of the robot by an interlocutor would be different, and the interaction style is assumed to be different.

Our experiments also did not consider dialogue context. As mentioned in Section 2.1, the interaction purpose and/or the shared planes play an important role especially in joint action [30]. Gorniak [45] focused on such dialogue context and studied situational representations used by humans while they are playing the game. In the cooperative task, we should consider such situational language uses by humans to provide further disambiguation. In our experiments, the robot used face direction and pointing gestures, and adjusted the amount of its speech and the gestures to clarify humans' referring behaviors by considering entrainment and its inhibition. Such adjustment becomes more difficult if a robot uses more variety of nonverbal behaviors. Therefore, in more complex situations, we should also use the situational context. For example, in a situation where humans often use gaze movement to convey their intentions, the robot should also prioritize the gaze information in its conversation strategy to effectively entrain the gaze behaviors.

To handle such a situational context, Mavridis and Roy [46] developed a grounded situation model (GSM) to ground situational representation in the visual scene. The robot which implemented the GSM reasoning resolved the situational context and created referring expressions [47]. Such symbol grounding related techniques would be helpful to treat situational contexts and reduce the ambiguity of human behaviors. We also note the research in the context of active perception, Mavridis and Dong [48] developed a conversational robot that was able to gather about its environment to resolve ambiguous verbal expressions through sensory actions and by asking its human partner questions. If we do not assume total sensory accessibility of the referents, the robot should implement the cost of its action in its conversation strategy to resolve the ambiguity of referring behaviors as well as entrainment phenomena.

There was a large difference in the rate of redundant reference between Experiment 1 (the rate of implicit entrainment in congestion placement is about 20%) and Experiment 2 (the rate of implicit entrainment is about 40%), which might be due to differences in participants' attributes between the experiments. The participants in Experiment 1 were averaged 23.3 years of age and the participants in Experiment 2 were averaged 36.5 years of age. The participants' averaged years of age in Experiment 2 was higher than in Experiment 1 and the rates of redundant reference in Experiment 2 was higher than in Experiment

1. This is an interesting difference because it suggests that there might be a new factor contributing to the occurrence of entrainment and such a point of human attributes need to be further studied in future work.

In this study, we utilized the lexical and gestural entrainment. In addition to these entrainment phenomena, recently the emotional entrainment has been demonstrated in human–robot interaction studies [49,50]. Hashimoto et al. [49] found that the synchronized emotional expressions of the robot with humans lead to humans' comfortable state. Costa et al. [50] reported the relationships between the facial expressions of the robot and the emotions of the listener in the storytelling context. The emotional entrainment could be applicable in the conversation strategy and also need to be further studied towards effective and more natural interactions.

8. Conclusions

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 strategy: lexical entrainment, gestural entrainment, and entrainment inhibition. Based on these phenomena, we designed robotic behavior policies which suggest that robots should provide the 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 strategy 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 entrainment. We experimentally compared two interactive strategies to determine which strategy improves the performance and which strategy makes better impressions on people. Even though the results indicated that the participants evaluated the impressions of conversations with the implicit entrainment strategy more highly, the recognition performances of the two strategies were not significantly different, indicating that the implicit entrainment strategy is better than the explicit request strategy for object reference conversations with people.

1. The proposed strategy implicitly elicits redundant references and improves the performance of the indicated object recognition.
2. Even though the proposed strategy forms better impressions than the other interactive strategy that explicitly requests clarifications when people refer to objects, the recognition performances of the two strategies are not significantly different.

Author Contributions: Conceptualization: T.I., M.S. and K.S.; methodology: M.K., T.I. and M.S.; software: M.K.; validation: T.I. and M.S.; formal analysis: M.K.; investigation: M.K., T.I. and M.S.; resources: T.I. and M.S.; data curation: M.K.; writing—original draft preparation: M.K.; writing—review and editing: M.K., T.I. and M.S.; visualization: M.K.; supervision: K.S.; project administration: T.I., M.S. and K.S.; funding acquisition: M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was supported in part by JSPS KAKENHI Grant Numbers JP19J01290 and JP20K19897.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of ATR (protocol code: 20-501-3, date of approval: 31 March 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hatori, J.; Kikuchi, Y.; Kobayashi, S.; Takahashi, K.; Tsuboi, Y.; Unno, Y.; Ko, W.; Tan, J. Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 3774–3781.
2. Nickel, K.; Stiefelhagen, R. Visual recognition of pointing gestures for human-robot interaction. *Image Vis. Comput.* **2007**, *25*, 1875–1884. [CrossRef]
3. Schauerte, B.; Fink, G.A. 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 on ICMI-MLMI '10 Association for Computing Machinery (ACM), Beijing, China, 8–12 November 2010; p. 6.
4. Iwahashi, N. 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, ROMAN 2003, Millbrae, CA, USA, 2 November 2003; pp. 385–390.
5. Furnas, G.W.; Landauer, T.K.; Gomez, L.M.; Dumais, S.T. The vocabulary problem in human-system communication. *Commun. ACM* **1987**, *30*, 964–971. [CrossRef]
6. Shinozawa, K.; Miyashita, T.; Kakio, M.; Hagita, N. User specification method and humanoid confirmation behavior. In Proceedings of the 2007 7th IEEE-RAS International Conference on Humanoid Robots, Pittsburgh, PA, USA, 29 November–1 December 2007; pp. 366–370.
7. Wu, E.; Han, Y.; Whitney, D.; Oberlin, J.; MacGlashan, J.; Tellex, S. Robotic Social Feedback for Object Specification. In Proceedings of the AAAI Fall Symposia, Providence, RI, USA, 12–14 November 2015.
8. Kuno, Y.; Sakata, K.; Kobayashi, Y. 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, Kyoto, Japan, 27 September–4 October 2009; pp. 2025–2031.
9. Whitney, D.; Rosen, E.; MacGlashan, J.; Wong, L.L.S.; Tellex, S. Reducing errors in object-fetching interactions through social feedback. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 1006–1013.
10. Garrod, S.; Anderson, A. Saying what you mean in dialogue: A study in conceptual and semantic coordination. *Cognition* **1987**, *27*, 181–218. [CrossRef]
11. Brennan, S.E. Lexical Entrainment in Spontaneous Dialog. *Proc. ISSD* **1996**, *96*, 41–44.
12. Brennan, S.E.; Clark, H.H. Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.* **1996**, *22*, 1482–1493. [CrossRef] [PubMed]
13. Branigan, H.P.; Pickering, M.J.; Cleland, A.A. Syntactic co-ordination in dialogue. *Cognition* **2000**, *75*, B13–B25. [CrossRef]
14. Scheflen, A.E. The Significance of Posture in Communication Systems†. *Psychiatry* **1964**, *27*, 316–331. [CrossRef] [PubMed]
15. Kendon, A. Movement coordination in social interaction: Some examples described. *Acta Psychol.* **1970**, *32*, 101–125. [CrossRef]
16. Bergmann, K.; Kopp, S. Gestural Alignment in Natural Dialogue. *Proc. Ann. Meet. Cogn. Sci. Soc.* **2012**, *34*, 1326–1331.
17. Levitan, R. Entrainment in Spoken Dialogue Systems: Adopting, Predicting and Influencing User Behavior. In Proceedings of the 2013 NAACL HLT Student Research Workshop, Atlanta, GA, USA, 13 June 2013; pp. 84–90.
18. Lopes, J.; Eskenazi, M.; Trancoso, I. From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Comput. Speech Lang.* **2015**, *31*, 87–112. [CrossRef]
19. Fandrianto, A.; Eskenazi, M. Prosodic Entrainment in an Information-Driven Dialog System. In Proceedings of the 13th Annual Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012.
20. Lopes, J.; Eskenazi, M.; Trancoso, I. Automated two-way entrainment to improve spoken dialog system performance. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8372–8376.
21. Kimoto, M.; Iio, T.; Shiomi, M.; Tanev, I.; Shimohara, K.; Hagita, N. 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), Kobe, Japan, 31 August–4 September 2015; pp. 337–342.
22. Kimoto, M.; Iio, T.; Shiomi, M.; Tanev, I.; Shimohara, K.; Hagita, N. Alignment Approach Comparison between Implicit and Explicit Suggestions in Object Reference Conversations. In Proceedings of the 4th International Conference on Human Agent Interaction, Association for Computing Machinery (ACM), Singapore, 4–7 October 2016; pp. 193–200.
23. Kimoto, M.; Iio, T.; Shiomi, M.; Tanev, I.; Shimohara, K.; Hagita, N. Robot Confirmation Behavior to Improve Object Reference Recognition. *J. Robot. Soc. Jpn.* **2017**, *35*, 681–692. [CrossRef]
24. Kimoto, M.; Iio, T.; Shiomi, M.; Tanev, I.; Shimohara, K.; Hagita, N. Conversation Strategy Comparison between Explicit Request and Implicit Alignment in Object Reference Conversation. *J. Robot. Soc. Jpn.* **2018**, *36*, 441–452. [CrossRef]
25. Branigan, H.P.; Pickering, M.J.; Pearson, J.; McLean, J.F. Linguistic alignment between people and computers. *J. Pragmat.* **2010**, *42*, 2355–2368. [CrossRef]
26. Brennan, S.E. Conversation with and through computers. *User Model. User-Adapt. Interact.* **1991**, *1*, 67–86. [CrossRef]
27. Iio, T.; Shiomi, M.; Shinozawa, K.; Shimohara, K.; Miki, M.; Hagita, N. Lexical Entrainment in Human Robot Interaction. *Int. J. Soc. Robot.* **2014**, *7*, 253–263. [CrossRef]

28. Brandstetter, J.; Beckner, C.; Sandoval, E.B.; Bartneck, C. Persistent Lexical Entrainment in HRI. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, Association for Computing Machinery (ACM), Vienna, Austria, 6–9 March 2017; pp. 63–72.
29. Clark, H.H.; Brennan, S.E. *Grounding in Communication*; American Psychological Association: Washington, DC, USA, 1991; pp. 127–149.
30. Garrod, S.; Pickering, M.J. Joint Action, Interactive Alignment and Dialog. *Top. Cogn. Sci.* **2009**, *1*, 292–304. [CrossRef] [PubMed]
31. Charny, J.E. Psychosomatic Manifestations of Rapport in Psychotherapy. *Psychosom. Med.* **1966**, *28*, 305–315. [CrossRef] [PubMed]
32. Breazeal, C. Regulation and Entrainment in Human—Robot Interaction. *Int. J. Robot. Res.* **2002**, *21*, 883–902. [CrossRef]
33. Ono, T.; Imai, M.; Ishiguro, H. A Model of Embodied Communications with Gestures between Human and Robots. *Proc. 23rd Ann. Meet. Cogn. Sci. Soc.* **2001**, *732–737*.
34. Iio, T.; Shiomi, M.; Shinozawa, K.; Akimoto, T.; Shimohara, K.; Hagita, N. Investigating Entrainment of People's Pointing Gestures by Robot's Gestures Using a WOZ Method. *Int. J. Soc. Robot.* **2011**, *3*, 405–414. [CrossRef]
35. Holler, J.; Wilkin, K. Co-Speech Gesture Mimicry in the Process of Collaborative Referring During Face-to-Face Dialogue. *J. Nonverbal Behav.* **2011**, *35*, 133–153. [CrossRef]
36. Kawai, H.; Toda, T.; Ni, J.; Tsuzaki, M.; Tokuda, K. XIMERA: A New TTS from ATR Based on Corpus-Based Technologies. In Proceedings of the 5th ISCA ITRW on Speech Synthesis, Pittsburgh, PA, USA, 14–16 June 2004; pp. 179–184.
37. Lee, A.; Kawahara, T. 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, Sapporo, Japan, 4–7 October 2009; pp. 131–137.
38. Howard, I.P.; Rogers, B.J. *Binocular Vision and Stereopsis*; Oxford University: Oxford, MS, USA, 1995; ISBN 195084764.
39. Sugiyama, O.; Kanda, T.; Imai, M.; Ishiguro, H.; Hagita, N.; Anzai, Y. Humanlike conversation with gestures and verbal cues based on a three-layer attention-drawing model. *Connect. Sci.* **2006**, *18*, 379–402. [CrossRef]
40. Ball, K.K.; Wadley, V.G.; Edwards, J.D. Advances in Technology Used to Assess and Retrain Older Drivers. *Gerontechnology* **2002**, *1*, 251–261. [CrossRef]
41. Sanders, A.F. Some Aspects of the Selective Process in the Functional Visual Field. *Ergonomics* **1970**, *13*, 101–117. [CrossRef] [PubMed]
42. Seya, Y.; Watanabe, K. Objective and Subjective Sizes of the Effective Visual Field during Game Playing Measured by the Gaze-contingent Window Method. *Int. J. Affect. Eng.* **2013**, *12*, 11–19. [CrossRef]
43. Iio, T.; Shiomi, M.; Shinozawa, K.; Miyashita, T.; Akimoto, T.; Hagita, N. Lexical entrainment in human-robot interaction: Can robots entrain human vocabulary? In Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, USA, 10–15 October 2009; pp. 3727–3734.
44. Dogan, F.I.; Kalkan, S.; Leite, I. Learning to Generate Unambiguous Spatial Referring Expressions for Real-World Environments. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 4992–4999.
45. Gorniak, P.J. The Affordance-Based Concept. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2005.
46. Mavridis, N.; Roy, D. Grounded Situation Models for Robots: Where words and percepts meet. In Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, 9–15 October 2006; pp. 4690–4697.
47. Mavridis, N. Grounded Situation Models for Situated Conversational Assistants. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2007.
48. Mavridis, N.; Dong, H. To ask or to sense? Planning to integrate speech and sensorimotor acts. In Proceedings of the 2012 IV International Congress on Ultra Modern Telecommunications and Control Systems, Sankt Petersburg, Russia, 3–5 October 2012; pp. 227–233.
49. Hashimoto, M.; Yamano, M.; Usui, T. Effects of emotional synchronization in human-robot KANSEI communications. In Proceedings of the RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication, Toyama, Japan, 27 September–2 October 2009; pp. 52–57.
50. Costa, S.; Brunete, A.; Bae, B.-C.; Mavridis, N. Emotional Storytelling Using Virtual and Robotic Agents. *Int. J. Hum. Robot.* **2018**, *15*. [CrossRef]