Providing Route Directions: Design of Robot's Utterance, Gesture, and Timing

Yusuke Okuno Takayuki Kanda Michita Imai Hiroshi Ishiguro Norihiro Hagita

ATR Intelligent Robotics and Communication Laboratory
2-2-2 Hikaridai, Keihanna Science City,
Kyoto, Japan

{okuno, kanda, michita, ishiguro, hagita}@atr.jp

ABSTRACT

Providing route directions is a complicated interaction. Utterances are combined with gestures and pronounced with appropriate timing. This study proposes a model for a robot that generates route directions by integrating three important crucial elements: utterances, gestures, and timing. Two research questions must be answered in this modeling process. First, is it useful to let robot perform gesture even though the information conveyed by the gesture is given by utterance as well? Second, is it useful to implement the timing at which humans speaks? Many previous studies about the natural behavior of computers and robots have learned from human speakers, such as gestures and speech timing. However, our approach is different from such previous studies. We emphasized the listener's perspective. Gestures were designed based on the usefulness, although we were influenced by the basic structure of human gestures. Timing was not based on how humans speak, but modeled from how they listen. The experimental result demonstrated the effectiveness of our approach, not only for task efficiency but also for perceived naturalness.

Categories and Subject Descriptors

H.5.2 **[Information Interfaces and Presentation]**: User Interfaces-Interaction styles

General Terms

Design, Experimentation, Human Factors.

Keywords

Route directions, Gesture, Timing

1. INTRODUCTION

Social robots are embodied to be physically co-located with people. A series of studies has demonstrated the importance of robots' physical existence [10, 24, 28]. In addition, social robots are often equipped with human-like body properties. Previous studies have demonstrated that their human-like body properties enable them to interact naturally with people (as discussed further in Section 2.2).

We believe that one promising application of social robots is to provide route directions. Figure 1 shows a possible scene of providing route directions: a person is glancing around and looking for a shop (Figure 1 left), and a robot approaches her, asks for her destination (center), and provides the directions (right). A robot has a number of appropriate features for this task since it is physically co-located with people, it can proactively approach a person who needs such information, and then provide it "naturally" with its human-like body properties.

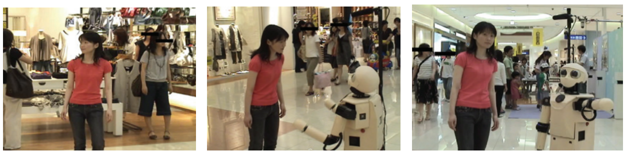


Figure 1. Robot gives route directions

What constitute good route directions from a robot? If the destination is within a visible distance, the answer might be intuitive: a robot would say "the shop is over there" and point. However, since the destination is often not visible, a robot needs to utter several sentences accompanied with gestures. We designed our robot's behavior to enable the listener to intuitively understand the information provided by the robot. This paper illustrates how we integrate three important factors – *utterances*, *gestures*, and *timing* – so that the robot can provide route directions.

2. RELATED WORKS

2.1 Utterance

With respect to human route direction behavior, Daniel *et al.* makes the following insightful observations: "A remarkable fact about route directions is that they do not always make it easy for people to reach their goal." They add that "ambiguous and confusing descriptions are known to be inefficient," and "descriptions that are too long and too detailed, however correct, become too difficult to memorize." They also established a "skeletal description," which consists of a series of sentences where each sentence contains a pair that consists of a landmark and an action [6]. We follow their study to optimize a robot's verbal information.

Also in robotics, there are studies related to route directions. One type is concerned with providing a route to a robot to navigate that robot, e.g., the study of hand-written maps [32]. Another type deals with tour guide robots that move among people to help them navigate (for example, see [4, 23]). In another study, a robot gives route directions [22]; but this previous study did not include landmarks in verbal information, since it was conducted in a simple corridor setting that did not require landmarks for route directions.

HRI'09, March 11–13, 2009, La Jolla, California, USA.

2.2 Gesture

2.2.1 Historical Discussion: performed for the speaker or listener?

In previous studies of human communication, research concentrated on why people perform gestures. Are gestures primarily for speakers or for listeners?

On one hand, some evidence suggests that a gesture is produced within a speaker's speaking process. For example, Krauss argued that gestures help people speak about spatial contents, because people became disfluent when they were inhibited from gesturing and people who gesture more speak more fluently [14]. Trafton *et al.* observed people's gestures to study their spatial working memory [35].

On the other hand, gestures have also been found to be directed toward listeners. Gesturing plays various roles in utterances [9]. Alibali reported that speakers who gesture intend to be communicative, and the produced gestures in fact help listeners understand spatial information, particularly when the message is complicated or unclear [1].

McNeill introduced the idea of a growth point, which suggests a human mechanism to simultaneous generate both utterances and gestures [16, 17]. He argues that the controversy about the benefit of gestures for speaker or listener is too simplistic and suggested that "every gesture is simultaneously 'for the speaker' and 'for the listener' [17]." In reality, it is impossible for humans to differentiate the role of their gestures as purely for speakers or for listeners.

2.2.2 Gestures in Route Directions

A couple of studies have been performed that are specific to route directions. Allen found that people performed deictic gesture (68%) more often than other gestures such as iconics (20%), beats (11%), and metaphorics (2%). This study also found that people who spoke fast do more gesturing, which implies that gestures make utterances fluid [2]. In a study of the natural behavior of an embodied conversation agent, human behavior in route directions was investigated. People usually embrace the route perspective (the perspective of the person following the route) (54.1%) rather than a survey perspective (16.3%) [13, 33].

Kita found that when a person utters a directional expression such as "turn right," he performs representational gesture to describe the action/direction of "right." The explaining person turns his torso orientation so that his "right" side matches with the direction of the route that will "turn right" [11]. In other words, this finding suggests that it is natural for a human speaker to coordinate his/her body orientation when describing a direction.

2.2.3 Gestures in Human-Robot Interaction

Joint attention and deictic gestures have frequently been used in HRI [19, 27, 34]. In addition to deictic gestures, humans also use their body to control the conversation flow by turn-taking, which is also replicated with robots, including gazing [3, 18, 30, 38], nodding [3, 21, 31], and arm movements [8].

Another study discussed the implementation of route directions in an embodied conversational agent [13]. The authors imitated human speaker's behavior including sentences and iconic gestures. This seems appropriate for creating the human-like behavior of virtual agents; but, their knowledge is probably not applicable to our robot, which must exploit its physical existence using deictic gestures more than iconic gestures to directly maps information into real spaces.

One previous study dealt with a robot that provided route directions. Ono *et al.* revealed that appropriate body orientation is important for conveying such directions as "right" and "left." They also suggested a possibility that gaze and body movements of the robot would reduce the time required for subjects to reach their destination, but yet they did not prove its contribution [22]. Our study aims to move one step beyond Ono's study to reveal how gestures contribute to the route directions.

2.3 Timing

We can categorize previous timing studies into two aspects. One is the timing between turns. Turn-taking [26] involves a pause between turns [15]. Jaffe *et al*. reported that in ordinary conversations the length of the pause ranged from 620 to 770 ms [7]. Nagaoka *et al.* revealed that a switching pause affected the impression of others [20]. Their study showed that Japanese people's switching pause averaged 590 ms.

Such a natural pause in human communication had been replicated in HRI. Yamamoto *et al.* compared a robot's response time in a greeting interaction and found that, after the user's action, a starting motion of about 300 ms and a starting utterance of about 600 ms were the most preferred timings. Shiwa *et al.* found that people preferred 1000 ms delayed response more than instant responses and also demonstrated that such conversational fillers as "*etto*" can help a robot comfortably placate a user when it cannot respond within a second [29]. Robins *et al*. explored how different response times change user reactions to a robot in a setting where a child and a robot are playing drums together [25].

The other aspect of previous studies is the timing between the robot utterance and its motion. For example, Yamamoto *et al.* found that when exchanging greetings, users prefer a robot that utters after 300 ms to its greeting motion [36]. Yamazaki *et al.* demonstrated the importance of the gaze timing, which should be placed at the relevant transition places for turn-taking [38].

In this paper, we explore a third timing aspect: the pause duration between the robot's utterances. This pause is resembles the one in the following passage: "Please go straight. (*pause*) Then you will see a post office." Unlike the switching pause in turn-taking, this pause should not cause the switch of speakers. In speech synthesis studies, vowel pause (typically 100-150 ms) and syllable pause (typically 200-300 ms) were modeled from human pauses (for example, see [39]). In contrast to such pauses between words within a sentence, pauses between sentences, which are highly variable and depend on speakers and their tasks, often ranging from 300 to 1000 ms [40, 41]. Pause durations also depend on cultures. For instance, Campione *et al.* revealed that the average duration varies from 400 to 550 ms among five languages [5]. The pause duration might be slightly shorter in Japanese, as Nagaoka *et al.* found by analyzing a large corpus of spoken Japanese. They reported that the majority of pauses between sentences of the same speaker are less than 400 ms [12].

Although these previous studies modeled the pause duration between sentences by modeling human speech, in this study we focus on modeling from the listener's perspective; to our knowledge, this is yet an underexplored aspect.

Note that both of the previous engineering realizations for route directions did not consider timing [13, 22].

3. Modeling of Robot's Route Directions

We modeled the generation process of a robot's route directions and divided it into three models: *utterances*, *gestures*, and *timing*. Figure 2 shows the information flow among these three models. First, a robot generates *utterance*. Second, it combines an *utterance* with a *gesture*. Finally, it expresses route directions with appropriate *timing*. For each model, we took different approaches, as summarized in Table 1.

Table 1. Modeling of route directions

Model	Design method
Utterance	Literature review
Gesture	Design consideration
Timing	Modeling from human behavior

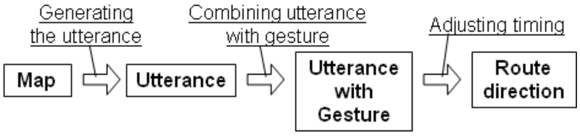


Figure 2. Generation process of robot's route directions

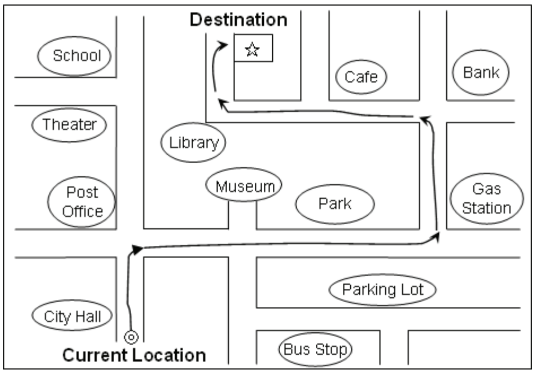


Figure 3. Map

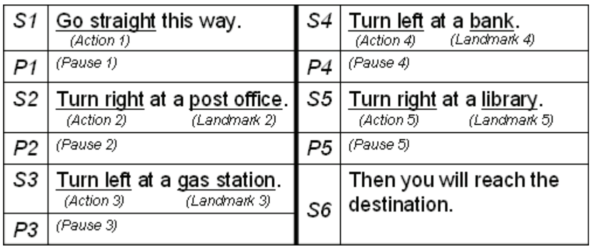
3.1 Utterance

Providing route directions is described as a process of providing knowledge about directions by means of utterances and gestures in combination. Usually, the source of the information is a geographical map (e.g., Figure 3) which contains locations of buildings and streets. A robot decides the way to the destination from the map, and then makes sentences to describe the way. For developing the utterance model, we did a literature review and decided to rely on the work conducted by Daniel *et al*. [6], who concluded that route description should contain minimal information that is neither too short to avoid ambiguity nor too detailed. From this standpoint, they proposed a "skeletal description" that contains minimal sets of information.

A "skeletal description" [6] consists of a series of sentences, and each sentence consists of a pair comprised of a landmark and an action. An action is an instruction about walking behavior, such as "go straight," "turn left," or "turn right." A landmark is an easy-to-find building in the environment where people are instructed to take an action, such as a bank, a post office or a library. Thus, a sentence should be something like, "turn left at a bank."

Following the "skeletal description," the robot uses these sentences to provide information about how to reach the destination. Table 2 shows an example of utterances for route directions between two places on the map shown in Figure 3. In the table, *S1-6* indicates sentences uttered by the robot, and *P1-5* indicates pauses between them.

Table 2. Utterance based on skeletal description



S1	Go straight this way. (Action 1)	S4	Turn left at a bank. (Action 4) (Landmark 4)
P1	(Pause 1)	P4	(Pause 4)
S2	Turn right at a post office. (Action 2) (Landmark 2)	S5	Turn right at a library. (Action 5) (Landmark 5)
P2	(Pause 2)	P5	(Pause 5)
S3	Turn left at a gas station. (Action 3) (Landmark 3)	S6	Then you will reach the destination.
P3	(Pause 3)		

3.2 Gesture

In route directions, since people also use gestures (see Section 2.2.2), we wonder what kind of robot gestures could help people understand route directions. From the literatures [2, 9, 11, 13, 22, 33], we retrieved gestures often used in route directions and made a list that contains the following four types: "deictic gesture," "orienting body direction," "iconic gesture (expressing landmarks)," and "beat gesture." Table 3 shows a summary of these gesture types. The following three aspects are considered.

Aspect 1: Is the gesture accompanied by speech?
Gestures are produced in a speaker's process for formulating speech [14]. All of the listed gestures are produced in the process of formulating speech, because they are used in the context of providing route directions.

Aspect 2: Does the gesture help human listeners understand?
Gestures help listeners understand utterances when a message is complicated or unclear (e.g., [1, 35]). Since our utterance contains sufficient information, the information that could be conveyed by gesture is also conveyed by utterances. Thus, gesture could be considered as a redundant message in our case (for example, pointing the "right" direction with saying "turn right").

Under this assumption, we considered that "deictic gestures" would be useful, as pointing direction could visually provide a clear supplement for such utterance "turn right" with its absolute direction to walk. This is the gesture that is most often used in route directions [2]. At least, there is little risk that it would cause misunderstanding.

"Orienting body direction" would be useful in cases where "right" and "left" are inconsistent between the speaker and the listener, e.g., when facing each other [22]; however, once they establish a direction, e.g., standing to the side, changing body orientation does not help much. On the other hand, one risk of using "orienting body direction" is that when the robot frequently orients its body orientation and if people do not follow, they lose the coordination of body orientation; thus, it is not clear whether this would be useful.

"Iconic gestures (expressing landmarks)" could be useful, though they depend on the available landmarks. Some landmarks are easily represented as icons, but many are difficult, such as gas stations, banks, or libraries. There could be a risk here: if people would not understand the iconic gesture of the robot, they might be confused about it.

"Beat gestures" are usually used for emphasizing important parts in speech; however, since "skeletal description" contains minimal information, they are not so useful in our case. Also, there could be a risk that people would not understand the meaning of beat gesture of the robot.

Aspect 3: Can robots express the gesture?
This aspect depends on the robot's shape. Robots have often limited degrees of freedoms in comparison to humans. In our case, the robot can perform deictic and beat gestures and orient its body direction, but it had difficulty performing iconic gestures well because they usually require hands and fingers to create shapes, which was impossible for our robot.

Table 3. Four types of gestures [2, 9, 11, 13, 22, 33]

	Deictic	Orienting body direction	Iconic (landmarks)	Beat
Useful for speakers	**yes**	yes	yes	yes
Useful for listeners	**definitely yes**	**Yes in the beginning**, unclear for later	maybe	no (unless utterance is redundant)
Robots can express well	**yes**	**yes**	difficult	maybe

Overall, we decided to use deictic gestures after the robot and listening person orient their body directions. We did not use iconic and beat gesture, because it was unclear whether it would work positively or negatively; we intended to reveal the usefulness of the most promising gesture, "deictic gesture," and placed the other gestures for a topic for the future study.

Figure 4 shows the deictic gesture we designed. The deictic gesture is used to point the absolute direction. (It is the iconic gesture that is used to represent the action of the movements). For example, for the case when the robot utters the following sentence: "Go straight (Figure 4 (a)) this way. Turn right (Figure 4 (b)) at a post office. Turn left (Figure 4 (a)) at a gas station." Since the deictic gesture points the absolute direction, it performs gesture (a) but not the gesture (c) for the last phrase "turn left."

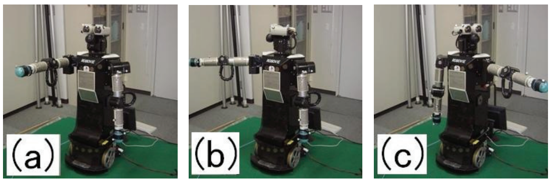


Figure 4. Deictic gestures

3.3 Timing

The robot pauses between sentences when it speaks (Figure 5). This model of *timing* decides the duration of these pauses. We take two different approaches for modeling the timing: from the speaker's perspective (i.e., natural timing for speaking) and from the listener's perspective (i.e. time needed to understand). Later we experimentally decided which one enables a robot to provide better route directions (The experiment is explained in Section 4).

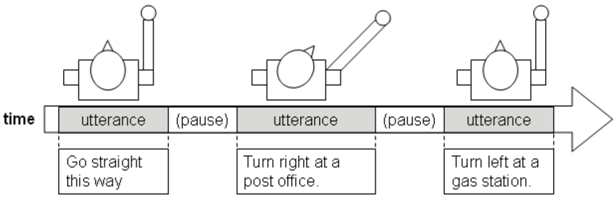


Figure 5. Pauses during route guidance

3.3.1 *Modeling from Speaker's Timing*

Modeling of human speaker behavior is commonly used for creating naturalness in computers, for example, speech synthesis or CG agent [12, 13, 18, 20, 21, 36, 37, 39, 40, 41]. With this modeling approach, the robot behaves similar to humans. Thus, a possible hypothesis is that a robot could naturally provide route directions based on the timing modeled from a human speaker's timing.

To model the timing of speakers, we measured their pause durations. We asked six students in our laboratory who did not know the purpose of the study to read the sentences generated with the "skeletal description" described in Section 3.1. The description consisted of six sentences. They looked at the description for a minute, and then read it naturally to a listener. The listener was standing in front of the speaker, but did not provide a particular reaction (e.g. nodding) to the speaker. We measured the pause duration between sentences.

Figure 6 shows the average duration of the pauses of the six speakers. *Pi* represents a pause after *i*-th sentence (the same symbols used in Table 2). The *P1* to *P5* average ranged between 0.42 and 0.66 [sec], which seems a reasonable value in comparison with previous speech synthesis studies [12, 40, 41].

Based on this "speaker" model, the robot used these measured *P1* to *P5* for the pause duration in route directions.

3.3.2 *Modeling from Listener's Timing*

An alternative timing approach is to model the time that people take to understand the utterances. Listening to route directions is a cognitively demanding task that requires the comprehension of spatial relationships and the memorization of routes and landmarks. Thus, making a robot who takes enough time after each sentence is a reasonable and could increase understanding in listeners.

We modeled the time that people take to understand route directions. For the data collection, a robot provided route directions to a human participant. We used the same robot as in the experiment reported in Section 4. The robot spoke the sentences generated based on the "skeletal description" described in Section 3.1. The description consists of six sentences. In the data collection, the robot uttered sentences one by one; participants were asked to verify that they understood the utterance after each sentence the robot uttered; after verification, the robot started the next sentence. In this data collection, the robot was controlled by wizard-of-oz method.

Eight university students participated in this experiment. We measured the time between the end of the robot's utterance and when participants reported that they understood. Each participant repeated this measurement four times for different routes. Figure 6 (upper line) shows the duration average required by listeners to understand the utterance. It is the average of eight participants with four trials. On average, *P1* to *P5* ranged between 1.03 and

3.07 [sec]. Based on this "listener" model, the robot uses these measured *P1* to *P5* for pause durations in the route directions. These values are quite long in comparison to those found in previous speech synthesis studies [12, 40, 41].

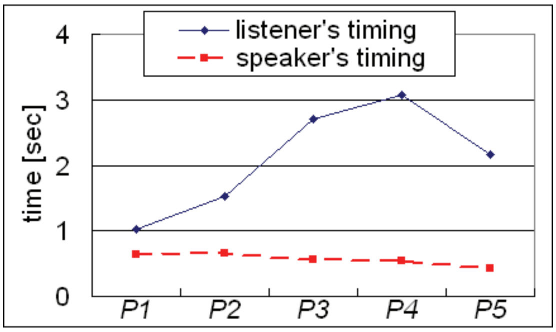


Figure 6. Timing model

4. EXPERIMENT

We conducted an experiment to measure an effect of gestures and to find a better model of timings. In addition, we compared the robot's task performance with human's route directions to evaluate the effectiveness of our model.

4.1 Method

Participants

21 undergraduate, native Japanese speakers (14 males and 7 females) participated in our experiment for which they were paid.

Settings

We used "Robovie," a 1.2-m tall communication robot with a 0.5-m radius whose human-like upper body is designed for communication with humans. It has a head (3 DOFs), eyes, and arms (4*2 DOF). A speaker was attached on its head. With its 4-DOF arms, it can perform deictic gestures (Figure 4).

Figure 7 shows the experimentation environment, which is a 3 x 3 m space separated from the rest of the room by partitions. An A0 size picture of a way in a town is presented on the face of one partition. A speaker (i.e. a human or a robot depending on condition) and a listener (participant) stood near the picture.

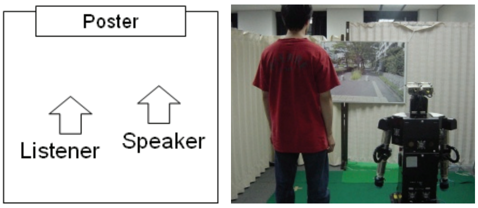


Figure 7. Experiment environment

Conditions

There were two conditions for both gestures and timing. In addition, two people provided route directions.

Conditions for the robot

The robot gave directions based on the description shown in Section 3.1. They were based on the "skeletal description," and each had six sentences. (Scene of the experiment is also available as the video attachment.)

a) Gesture

With: The robot performed the deictic gestures described in Section 3.2 (Figure 4).

Without: The robot did not do any gestures. Thus, no arm movements were expressed.

b) Timing

Speaker: The robot uttered sentences based on the speaker-based model described in Section 3.3.1.

Listener: The robot uttered sentences based on the listener-based model described in Section 3.3.2.

Condition with a Human Speaker

Two human speakers gave route directions. They were from our laboratories but did not know the purpose of the experiment. They received a map (e.g. Figure 3) 15 minutes before the start of the experiment and were told to provide a route to the destination. Since we did not allow them to use such ordinals as the "n-th corner," they needed to use landmarks and actions. We did not provide any further instruction. Thus, some performed detailed route directions, and some gave directions that resembled the skeletal descriptions.

Procedure

The experiment was a within-subject design. Participants repeated the session for all conditions. The order of sessions was counterbalanced. The given route and the destination were different every time. They received a 10-minute break between each session.

At the first session, participants were instructed to imagine getting lost at an unfamiliar place on their way to a famous restaurant. They were instructed to ask about a route, to learn it, and then draw it on a piece of paper after listening to the route directions. They were also told that they had to learn both landmarks and actions to reach the destination, since the map was not on a grid. Participants were positioned at the "listener" position in Figure 7, and the robot/human speaker stood at the "speaker" position.

At each session, the participants listened to the route directions. After the route directions had been provided, they drew a map and completed questionnaires. Note that we prohibited participants from asking for the route directions again; thus, the route directions were given only once per condition.

4.2 Measurement

We conducted two types of evaluations. One was about the task performance. That is, how well participants understood the route directions.

Correctness: We asked participants to draw a map of the route that was explained during the experiment. We counted the number of correct actions and landmarks on the maps. Each description had 4 landmarks and 4 actions; we excluded the first sentence from the score (*S1* in Table 2) because it was always "go straight this street" and all participants answered it correctly. The score is ranged from 0 to 8. For example, three correct landmark and two correct actions were scored as 5.

The participants were also asked to provide comments about their impressions with respect to the following items on a 1-to-7 scale where 1 stands for the lowest evaluation and 7 for the highest.

Easiness: How easy/difficult was it to understand the route?
Naturalness: How natural/unnatural were the route directions?

4.3 Hypothesis and Predictions

Since previous studies indicated the usefulness of human gestures for human listeners, we thought that the robot's gestures would help listeners understand route directions, even when the gesture conveyed a redundant message represented in parallel by utterances, such as pointing "right" direction while saying "turn right".

Since Figure 6 shows that listeners need more time to understand route directions than the pause duration of speakers, we thought that a speaker-based model would not offer enough time for listeners to understand.

Based on these considerations, we made the following predictions:

1. When participants listen to the route directions with gestures, the correctness scores and easiness ratings will outperform the cases when they listen without gesture.
2. When participants listen to the route directions with the timing of the listener-based model, correctness scores and easiness ratings will outperform the cases when they listen to the route directions with the timing of the speaker-based model.

5. RESULTS

5.1 Verification of Predictions

For the correctness score results (Figure 8), a two-way repeated-measures Analysis of Variance (ANOVA) was conducted with two within-subject factors, *gesture* and *timing*. A significant main effect was revealed in both the gesture factor (F(1,20)=16.055, p<.005) and the timing factor (F(1,20)=6.757, p<.05), but no significance was found in the interaction within these factors (F(1,20)=.323, p=.576).

For the results of the easiness ratings (Figure 9 (a)), a two-way repeated-measures ANOVA with two within-subject factors, *gesture* and *timing*, was conducted. A significant main effect was revealed in both the gesture factor (F(1,20)=13.945, p<.005) and the timing factor (F(1,20)=12.105, p<.005). Interaction within these factors was also significant (F(1,20)=6.448, p<.05).

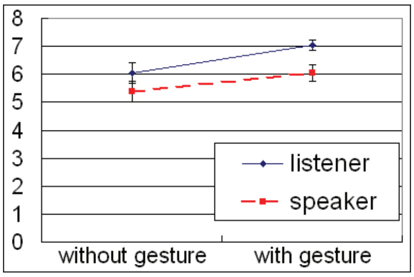
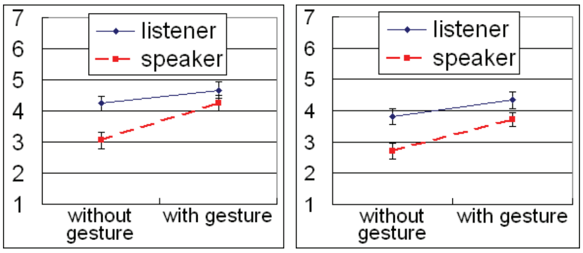


Figure 8. Experimental results of correctness scores

This interaction suggests that each of factors contributed to the easiness but the combination of two factors (with-gesture and listener-based model) did not improve the easiness twice. There was a simple main effect in the gestures in the speaker-based model (F(1,20)= 28.027, p<.001), while no significant main effect was observed in the gestures in the listener-based model (F(1,20)=2.077, p=.165). There was a simple main effect in the timing in the without-gesture condition (F(1,20)= 30.941, p<.001), but no significant main effect was observed in the timing in the with-gesture condition (F(1,20)=1.709, p=.206).

Overall, predictions 1 and 2 are supported.



(a) easiness (b) naturalness

Figure 9. Experimental results of subjective evaluations

5.2 Comparison of Naturalness Ratings

A two-way repeated-measures ANOVA was conducted with two within-subject factors, *gesture* and *timing*, for the result of naturalness (Figure 9 (b)). A significant main effect was revealed in both the gesture factor (F(1,20)=8.928, p<.01) and the timing factor (F(1,20)=12.308, p<.005). Interaction within these factors was also significant (F(1,20)=5.525, p<.05). We analyzed the simple main effects in the interaction within these factors. It was almost significant in the gestures in the listener-based model (F(1,20)=3.467, p=.077), and significant in the gestures in the speaker-based model (F(1,20)=14.000, p<.005), in timing in the with-gesture condition (F(1,20)=5.972, p<.05) as well as the without-gesture condition (F(1,20)= 15.838, p<.005). Overall, participants rated the naturalness ratings higher for the robot that uttered with gestures and with the timing of the listener-based model. The combination of the two factors did not improve the naturalness rating twice, as in the case of the easiness ratings.

5.3 Comparison with the Route Directions Given by Humans

To analyze how well the robot achieved in route directions, we compared its route directions and humans' (Figure 10). The robot condition is the one "with gesture" and "listener-based model" condition, which was found to be the best design of the robot's route directions.

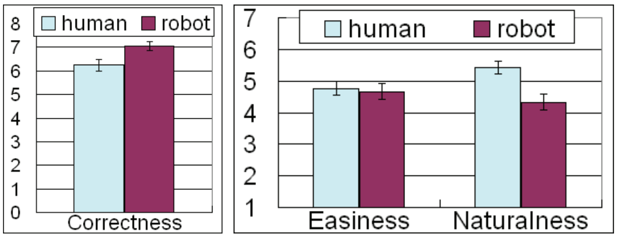


Figure 10. Comparing route directions of human and robot

A one-way repeated-measures ANOVA was conducted with one within-subject factor, *speaker* (human or robot). A robot performed better in correctness scores (F(1,41)=7.920, p<.01), no significant difference in easiness ratings (F(1,41)=.067, p=.796), and humans gave better impression in naturalness ratings (F(1,41)=10.348, p<.005).

It is interesting that the robot provided route directions led to better correctness scores; however, the participants perceived it as not differently easy, and rather unnatural than the one provided with robots. This seems to suggest that the robot's route directions could be improved. For example, its utterance is too simple in contrast to humans' utterances. They used more conversational fillers and conjunctions. In addition, it could be not so simple to compare such naturalness, as people might apply different criteria of naturalness toward the robot and humans.

6. DISCUSSION

6.1 Summary

This study reports the design of route directions provided by a humanoid robot, which based on three models: utterance, gesture, and timing. The findings from the experiment can be summarized as follows:

(1) The effect of the robot's gestures is demonstrated in route directions where deictic gestures are used in parallel with utterances. That is, even though utterances provide enough information for a listener to understand the way, the gestures add information to the utterances.

(2) The importance of timing is highlighted. People prefer a pause duration modeled from time required to understand, even though the pause duration far exceeds common pause timing. Such long pause duration could be interpreted as unnatural, but interestingly the experimental result revealed that the participants rated it as more natural; our interpretation is that participants were busy understanding route directions so that they did not perceive unnaturalness in such a long silence.

6.2 Design Implication

We believe that this study shows one robot design approach that naturally and effectively performs interactions. Even with our robot's limited capabilities, we can still learn from humans' behavior (e.g. gestures) to improve our robot. Moreover, robots can be designed to perform such non-verbal behaviors as gestures and timing just for the sake of listeners.

The comparison of the robot and humans suggested important insights into naturalness. As the result of these designs, the robot performed better than the humans in terms of the task performance (i.e. correctness of the listening to the person's memory of the way); among the comparison of factors of gesture and timing, this design is evaluated as the best in terms of both correctness and naturalness. However, when we compare this with human's route directions, people still evaluate humans' route directions as significantly more natural. Although there is much to be improved within the route directions situation, perhaps, it might also need a long way to go toward letting people perceive that they are "interacting naturally" with robots; for example, it would be also affected by the fact that the robot is not natural existence but artificial existence.

6.3 Limitations

This study is one of the first systematic studies of robot's route directions, thus we focused on the primary aspects: effectiveness of gestures and modeling of listener-based timing, whereas there are many factors not yet explored. Thus, the generality of the findings remains limited. We hope that the rest of the factors can be explored later.

First, the applicability of the gesture model is limited to the robot used in this study. However, Robovie's arm is an ordinal one, which is similar to other humanoid robots', such as ASIMO (Honda Motors). Thus, the gesture performed by Robovie could be similarly performed by these robots; and we believe that the same gesture effect would occur with these robots. In this study we only tested the effectiveness of deictic gestures, thus the effect of gestures could be improved if we use other types of gestures. In particular, it would be possible for more complicated robots with higher DOFs in their arms such as HRP-3 (Kawada Industry), to perform other gestures such as iconic gestures, which would contribute even better to route directions.

The timing model depends on the number of sentences and complicity of contents. It also depends on the capability of speech synthesis. When we interviewed to the participants, some of them mentioned about the difficulty to listen to the robot's utterances due to its unusual accent and intonation. Technologies are advancing year by years, thus, in the future it would be possible for people to understand what the robot speaks in shorter time than what it was possible in this study. This aspect also depends on languages. That is, the finding is limited to Japanese.

The comparison with humans only demonstrates the performance without any exchange of turns. To match with the robot's capability we inhibited people from asking again. However, humans interact more flexibly. In human route directions, it is often observed that a listening person asks the speaking person to repeat again. Thus, the result of this study does not demonstrate that robots are in general better than humans in route directions.

The experiment was conducted as in within-subject design; it is controversial in terms of reliability, since within-subject design would suffer from carryover effects while between-subject design requires a large amount of participants to eliminate individual differences. We believe that the task in our experiment much depends on individual capability, such as understanding spatial relationships, thus within-subject is a better design for effective study.

7. ACKNOWLEDGMENTS

This research was supported by the Ministry of Internal Affairs and Communications of Japan.

8. REFERENCES

[1] Alibali, M. 2005. Gesture in Spatial Cognition: Expressing, Communicating, and Thinking About Spatial Information. Spatial Cognition and Computation. 5, 4, 307-331.

[2] Allen, G. 2003. Gestures Accompanying Verbal Route Directions: Do They Point to a New Avenue for Examining Spatial Representations? Spatial cognition and computation. 3, 4, 259-268.

[3] Breazeal, C., et al. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2005). 383-388.

[4] Burgard, W., et al. 1998. The Interactive Museum Tour-Guide Robot. In Proc. the National Conference on Artificial Intelligence (AAAI'98). 11-18.

[5] Campione, E., et al. 2002. A Large-Scale Multilingual Study of Silent Pause Duration. Speech Prosody. 199-202.

[6] Daniel, M., Tom, A., Manghi, E., Denis, M. 2003. Testing the Value of Route Directions through Navigational Performance. Spatial cognition and computation. 3, 4, 269-289.

[7] Jaffe, J., and Feldstein, S. 1970. Rhythms of dialogue. Academic Press. New York.

[8] Kanda, T., Kamasima, M., Imai, M., Ono, T., Sakamoto, D., Ishiguro H., and Anzai, Y. 2007. A humanoid robot that pretends to listen to route guidance from a human. Autonomous Robots. 22, 1, 87-100.

[9] Kendon, A. 2004. Gesture: Visible Action as Utterance.

[10] Kidd, C.D., and Breazeal, C. 2004. Effect of a Robot on User Perceptions. Int. Conf. on Intelligent Robots and Systems (IROS'04). 4, 28, 3559-3564.

[11] Kita, S. 2003. Interplay of gaze, hand, torso orientation and language in pointing. In Kita, S. (Ed.) Pointing: Where Language, Culture, and Cognition Meet.

[12] Komori, M., Yamamoto, Y., Nagaoka, C. 2006. Manipulation of pause durations for the facilitation of understanding of speech played back at higher rate. The Japanese journal of ergonomics. 42, 2, 64-69. (in Japanese)

[13] Kopp, S., Tepper, P.A., Ferriman, K., Striegnitz, K., and Cassell, J. 2008. Trading Spaces: How Humans and Humanoids use Speech and Gesture to Give Directions. In Nishida, T. (Ed.) Conversational Informatics: An Engineering Approach.

[14] Krauss, R.M. 1998. Why do we gesture when we speak? Current Directions in Psychological Science. 7, 54-59.

[15] McLaughlin, M.L. 1984. Conversation: How talk is organized.

[16] McNeill, D. 1987. Psycholinguistics: a new approach.

[17] McNeill, D. 2005. Gesture and thought.

[18] Mutlu, B., Hodgins, J.K., and Forlizzi, J. 2006. A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior. HUMANOIDS'06. 518-523.

[19] Nagai, Y., et al. 2003. Emergence of Joint Attention based on Visual Attention and Self Learning. Int. Symposium on Adaptive Motion of Animals and Machines.

[20] Nagaoka, C., et al. 2005. Influence of Response Latencies on Impression Evaluation of Speakers in Dialogues. Technical Report of IEICE. 104, 745, 57-60. (in Japanese)

[21] Ogawa H., and Watanabe, T. 2001. InterRobot: speech-driven embodied interaction robot. Advanced Robotics. 15, 3, 371-377.

[22] Ono, T., Imai, M., Ishiguro, H. 2001. A Model of Embodied Communications with Gestures between Humans and Robots. Proc. Annual Meeting of the Cognitive Science Society (CogSci2001). 732-737.

[23] Pacchierotti, E., Christensen, H.I., Jensfelt, P. 2006. Design of an office guide robot for social interaction studies. Int. Conf. Intelligent Robots and Systems (IROS2006), 4965-4970.

[24] Powers, A., Kiesler, S., Fussell, S., Torrey C. 2007. Comparing a Computer Agent with a Humanoid Robot. ACM/IEEE Conf. on Human-Robot Interaction (HRI2007). 145-152.

[25] Robins, B., Dautenhahn, K., Boekhorst, R.te, and Nehaniv, C. L. 2008. Behaviour Delay and Robot Expressiveness in Child-Robot Interactions: A User Study on Interaction Kinetics. ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI2008). 17-24.

[26] Sacks, H., Schegloff, E.A., and Jefferson, G. 1974. A simplest systematic for the organization of turn-taking for conversation. Language. 50, 4, 696-735.

[27] Scassellati, B. 2000. Investigating Models of Social Development Using a Humanoid Robot. Biorobotics. MIT Press.

[28] Shinozawa, K., et al. 2005. Differences in Effect of Robot and Screen Agent Recommendations on Human Decision-Making. Int. J. of Human-Computer Studies. 62, 267-279.

[29] Shiwa, T., Kanda, T., Imai, M., Ishiguro, H., Hagita, N. 2008. How Quickly Should Communication Robots Respond? ACM/IEEE Conf. on Human-Robot Interaction (HRI2008). 153-160.

[30] Sidner, A.L., Kidd, C.D., Lee, C., and Lesh, N. 2004. Where to look: a study of human-robot engagement. Intelligent User Interfaces (IUI'04). 78-84.

[31] Sidner, C.L., et al. 2006. The effect of head-nod recognition in human-robot conversation. ACM/IEEE Conf. on Human-Robot Interaction (HRI2006). 290-296.

[32] Skubic, M., Blisard, S., Bailey, C., Adams, J.A., Matsakis, P. 2004. Qualitative analysis of sketched route maps: translating a sketch into linguistic descriptions. IEEE Trans. on Systems, Man, and Cybernetics. Part B, 34, 2, 1275-1282.

[33] Striegnitz, K., Tepper, P., Lovett, A., Cassell, J. 2005. Knowledge representation for generating locating gestures in route directions. WS in Spatial Language and Dialogue.

[34] Trafton, J.G., Cassimatis, N.L., Bugajska, M.D., Brock, D.P., Mintz, F.E., Schultz, A.C. 2005. Enabling Effective Human–Robot Interaction Using Perspective-Taking in Robots. IEEE Transactions on Systems, Man and Cybernetics. Part A, 35, 4, 460- 470.

[35] Trafton, J.G., et al. 2006. The Relationship Between Spatial Transformations and Iconic Gestures. Spatial Cognition and Computation. 6, 1, 1-29.

[36] Yamamoto, M., and Watanabe, T. 2004. Timing control effects of utterance to communicative actions on embodied interaction with a robot. IEEE Int. Workshop on Robot and Human Communication (ROMAN2004). 467-472.

[37] Yamamoto, M., Watanabe, T. 2006. Time Lag Effects of Utterance to Communicative Actions on CG Character-Human Greeting Interaction (ROMAN2006). 629-634.

[38] Yamazaki, A., Yamazaki, K., Kuno, Y., Burdelski, M., Kawashima, M., and Kuzuoka, H. 2008. Precision timing in human-robot interaction: coordination of head movement and utterance. CHI '08. 131-140.

[39] Zellner, B. 1994. Pauses and the temporal structure of speech. In Keller, E. (Ed.) Fundamentals of speech synthesis and speech recognition. 41-61.

[40] Zvonik, E., Cummins, F. 2002. Pause duration and variability in read texts. Int. Conf. on Spoken Language Processing (ICSLP-2002). 1109-1112

[41] Zvonik, E., Cummins, F. 2003. The effect of surrounding phrase lengths on pause duration. European Conference on Speech Communication and Technology (EUROSPEECH-2003). 777-780.