A Model of Embodied Communications with Gestures between Humans and Robots

Tetsuo Ono (tono@mic.atr.co.jp)
ATR Media Integration & Communications Research Laboratories
2-2 Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0288 Japan

Michita Imai (michita@mic.atr.co.jp)
ATR Media Integration & Communications Research Laboratories

Hiroshi Ishiguro (ishiguro@sys.wakayama-u.ac.jp)
Faculty of Systems Engineering, Wakayama University

Abstract

In this paper, we propose a model of embodied communications focusing on body movements. Moreover, we explore the validity of the model through psychological experiments on human-human and human-robot communications involving giving/receiving route directions. The proposed model emphasizes that, in order to achieve smooth communications, it is important for a *relationship* to emerge from a mutually entrained gesture and for a *joint viewpoint* to be obtained by this relationship. The experiments investigated the correlations between body movements and utterance understanding in order to confirm the importance of the two points described above. We use robots so that we can control parameters in experiments and discuss the issues related to the interaction between humans and artifacts. Results supported the validity of the proposed model: in the case of human-human communications, subjects could communicate smoothly when the relationship emerged from the mutually entrained gesture and the joint viewpoint was obtained; in the case of human-robot communications, subjects could understand the robot's utterances under the same conditions but not when the robot's gestures were restricted.

Introduction

Why do people use gestures when communicating? A common scene on a street involving giving/receiving route directions is some person and a stranger making gestures together as if dancing synchronously and rhythmically (see Figure 2). These gestures appear not only when the person describes turns at visible locations, but also at invisible ones. Moreover, it has been shown that people are unable to achieve smooth communications if they are restricted from using spontaneous gestures (Ono et al., 2001). Consequently, gestures play an important role in human-human communications.

In this paper, however, we do not discuss *emblem gestures* such as the OK sign; these gestures have arbitrarily defined meanings and figures in social conventions. The target of our research is *mutually entrained gestures*, where a speaker and a hearer spontaneously and synchronously move their bodies according to the entrainment resulting from mutual actions and utterances. We focus on such gestures because smooth communications between humans can be expected when these gestures are used, as illustrated by the above example involving giving/receiving route directions.

In order to investigate the mechanism of the mutually entrained gestures described above, we conduct experiments on human-robot communications as well as human-human communications. The reason why we use a robot is that we can unrestrictedly design experiments by using a programmable robot's gestures. Moreover, an investigation of human-robot communications can contribute to research on the methodology of robot design and the interaction between humans and artifacts.

The purpose of this paper is to propose a model of embodied communications that can give an explanation for the mechanism of communications described above and, moreover, provide evidence for the validity of the model through psychological experiments. The main characteristic of our model is to focus on the *relationship* emerging from a mutually entrained gesture and the *joint viewpoint* obtained by the relationship. The experiments concretely investigate the correlations between body movements and utterance understanding in human-human and human-robot communications involving giving/receiving route directions. In such a task, it is hard for a person and a stranger to communicate with each other if they do not share the same viewpoint. Here, in order to obtain a joint viewpoint, both sides need to construct a relationship emerging from mutually entrained body movements. We investigate the process of communications in the experiments by using our implemented robot.

Embodied Communications

Previous Research on Gestures

Research on gestures conducted to investigate the mechanism of communications emerged around 1980. In this field, McNeil was the first to carry out cutting-edge research. He pointed out that gestures are synchronized with speech in communications, and thus both are closely connected in the cognitive system (McNeil, 1987). McNeil's research provided findings leading to the development of research on the functions of gestures.

However, previous research has mainly analyzed the speaker's gestures in communications. In other words, many researchers have analytically investigated the correlations between speech and the speaker's body movements. Consequently, their aim has been to explain an internal mechanism of an individual speaker. However,

these research works have not looked into the dynamical mutual interaction between a speaker and a hearer.

In contrast, we focus on the *dynamical mutual interaction* in human-human and human-robot communications involving giving/receiving route directions. Especially, the reason why we come to adopt this route directions is that spontaneous gestures such as pointing easily appear in this context. Kita (2000) analyzed a speaker's gestures for this task but did not deal with the dynamical mutual interaction between them. In this research, we are able to give evidence for a hypothesis in detail because we can control the parameters in experiments by using a robot.

Model of Embodied Communications

In this paper, we propose a model of embodied communications focusing on entrained body movements. Our model is basically described by the following formula:

$$\bigoplus(S,U) \rightarrow I$$

Here, $\bigoplus$ is a viewpoint for understanding an utterance in a situation, S is the situation around a speaker and a hearer, U is an utterance from the speaker, and I is information obtained by having understood utterance U.

For example, let us suppose that in a situation S' involving giving/receiving route directions, a person A utters "Go right" to a stranger B while both are facing each other. Let us further suppose that B understands from his/her viewpoint of A that the utterance means the "right" of A. In this case, the relation among the viewpoint, situation, utterance, and information is expressed as $\bigoplus_A(S',U_A) \rightarrow I_B$. However, B may instead understand from his/her viewpoint of himself/herself that the utterance means the "right" of him/her. In this case, the relation is expressed as $\bigoplus_B(S',U_A) \rightarrow I_B$.

The above ambiguity can be effectively solved by using an absolute coordinate system. For example, a person can clearly direct a stranger to a destination when both sides can use a visible landmark or object, or when both sides can construct a similar cognitive map (this assumes the stranger has previously visited the area). In this case, the viewpoint $\bigoplus$ is determined definitely.

However, a person cannot use an absolute coordinate system when landmarks and turns to the destination are invisible, or when a stranger has not visited the area before. In this case, it is difficult to maintain a joint viewpoint $\widehat{\bigoplus}$ in communications because the stranger is unable to imagine the route map of the person; the person's memory access also becomes overloaded.

As described in the Introduction, people seem to solve the problem of deciding a viewpoint by mutually entrained body movements. In other words, people first construct a relationship that emerges from a mutually entrained gesture. This relationship allows people to obtain a joint viewpoint. Finally, they can communicate with each other smoothly because of the utterance understanding achieved as a result of this joint viewpoint.

In our proposed model, the characteristic of communications discussed above is expressed as follows:

$$\widehat{\bigoplus}(\emptyset,U) \rightarrow I \quad (1)$$

$$O({}_iR_j) \rightarrow \widehat{\bigoplus} \quad (2)$$

$$E(torso,arms,eyes) \rightarrow {}_iR_j \quad (3)$$

Here, $\emptyset$ indicates the situation where there is nothing to point out, ${}_iR_j$ is the relationship between persons i and j, and O is a function for obtaining the viewpoint from the relationship. Moreover, *torso* and *arms* are expressions for entrained movements of the torso and arms, while *eyes* expresses the eye contact in communications. E is a function of the relationship emerging from the entrained movements.

These formulae express the process of communications involving giving/receiving route directions as follows. People cannot adopt an absolute coordinate system when they do not have a landmark or object to point out. Consequently, it is hard for them to achieve utterance understanding because of the difficulty of obtaining a joint viewpoint (Formula 1). To overcome this problem, they try to construct a relationship to obtain the joint viewpoint (Formula 2). This relationship emerges from mutual entrained body movements (Formula 3). Smooth communications can be achieved through these processes because the joint viewpoint makes both the speaker's utterance and the hearer's understanding easier.

In our model, we formalize the process of communications described above. We carry out psychological experiments to explore the validity of the model in the following two chapters.

Human-Human Communications

Experiments

Experiments on human-human communications were conducted by the following method.

Outline of experiments We focused on the interaction between a subject and a person involving giving/receiving route directions as an informant just happened to be passing by. Here, we investigated the appearance of their gestures, gestural arrangements, utterances, and the level of utterance understanding.

Subjects Ten undergraduate and graduate students (male and female). The subjects had not previously visited the experimental environment, and thus did not know the route to any destination at all.

Environment Figure 1 shows an outline of the experimental setup. These experiments were done in the hallways of a laboratory. Point A denotes the place where the route directions were given, and B and C denote the goals, i.e., a cafeteria and an information desk, respectively. Point T1 denotes a turn in the route from A to B, and Points T2-T4 denote turns from A to C. Only the corner of T1 is visible from A.

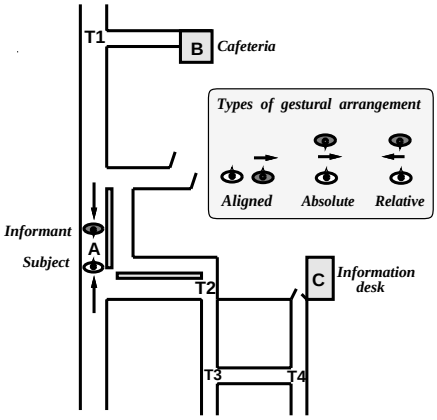

Figure 1: Experimental setup: arrangement of subject, informant, destinations, and turns.

Procedure The subjects received the following instructions from the experimenter (position A): "Ask a passerby the way to the cafeteria and the information desk and go to each place by yourself." The behaviors and utterances of the subjects and the persons were recorded with a camera and a microphone.

Evaluation The results of the experiments were evaluated from the record of the subjects' and the persons' behaviors, i.e., gestural arrangements, arm and elbow movements, and eye contact. In addition, we evaluated the time needed to communicate and the accuracy of the communicated information.

We classified the gestural expressions and arrangements into three categories, i.e., *aligned*, *absolute*, and *relative* gestures, following the literature (Kita, 2000) (see the upper right-hand side of Figure 1). To illustrate, let us assume that, at position A in Figure 1, an informant directs a subject to destination B by telling him/her to turn right at corner T1. In *aligned gesture*, an informant makes gestures to indicate his/her right aligning his/her torso orientation with that of the subject. In *absolute gesture*, an informant makes gestures to indicate the subject's right while facing the subject. In *relative gesture*, an informant makes gestures to indicate his/her right while facing the subject.

Results

First, the gestural expressions and arrangements that the subjects and the persons took were *aligned gesture* in nine out of ten cases and *relative gesture* in the remaining one case. These results were the same for both destinations. Next, Table 1 shows the analyzed results of synchronized gestures between the subjects and the persons. Synchronized arm gestures were observed in six out of ten cases in the route directions to destination B and eight out of ten cases to destination C. Here, synchronized arm gestures mean that the subject synchronously makes similar movements to the person's spontaneous arm movements (see Figure 2). In this experiment, all of the persons made arm movements. In addition, all of those who made synchronized gestures moved their extended arm right and left. Moreover, in all cases, they made eye contact. In particular, in the more complicated route to C, eye contact was made with high frequency.

Table 1: Results of entrained actions of arms and eyes in human-human interaction.

	Arm synchronized	Elbow extended	Eye contact (total)
Cafeteria	6/ 10	6/ 6	12 times/ 123 sec
Information	8/ 10	8/ 8	25 times/ 216 sec

Figure 2: Photo of mutually entrained gesture in human-human communications in the route direction.

Furthermore, the time needed to communicate was 17.2 seconds in the case of destination B but 32.2 seconds in the case of destination C. That is, the more complicated route direction statistically needed much more time ($t_{(18)} = 2.122,\ p < .05$). However, there was not much difference in the kinds of expressions used in the utterances between the two destinations. Eventually, all of the subjects could arrive at the two destinations. In other words, information was accurately communicated from the person to the subjects.

A summary of the experimental results is as follows. First, the persons acting as informants made spontaneous gestures not only when they described turns at visible locations but also at invisible ones. The subjects involuntarily made entrained and synchronized gestures to the persons.

We can assume the following relation between the experimental results and our proposed model. The subjects had not previously visited the experimental environment. Therefore, it was hard for the subjects to understand the persons' utterances because of the difficulty of obtaining a joint viewpoint (Formula 1). To overcome this problem, they tried to construct a relationship to obtain the joint viewpoint (Formula 2). This relationship emerged from mutually entrained body movements (Formula 3).

In the next chapter, we describe experiments on human-robot communications in order to investigate these mechanisms in detail. We can unrestrictedly design the experiments by using a programmable robot's gestures. Moreover, the investigation of human-robot communications can contribute to research on robot design and the interaction between humans and artifacts.

Human-Robot Communications

Experiments

Experiments on human-robot communications were conducted by the following method.

Outline of experiments We focused on the interaction between a subject and a robot as an informant involving giving/receiving route directions. Here, we investigated the appearance of the subject's gesture and the level of utterance understanding while changing the robot's gesture.

Subjects Thirty undergraduate and graduate students (male and female). The subjects were randomly divided into six groups. The subjects had not previously visited this experimental environment, as in the human-human experiments.

Robot Our robot system can make gestures by using the upper part of its body in the same way as a human (see Figure 3). The robot has two arms, two eyes, a mobile platform, and various actuators and sensors. With this equipment, the robot can generate almost all of the behaviors needed for communications with humans.

Environment Figure 4 shows an outline of the experimental setup. These experiments were done in the hallways and lobby of a laboratory. Points S and R denote the initial positions of the subject and robot, respectively. Point A denotes the place where the route directions were given, and B denotes the goal, i.e., the lobby. Points T1-T4 denote turns in the route from A to B, directed by the robot. Only the corner of T1 is visible from A.

Procedure The experiments consisted of the following six phases.

1. The subjects received the following instructions from the experimenter (position S): "Ask the robot the way to a lobby and go there by yourself." The question to the robot was specified as follows: "Tell me the way to the lobby."

2. The subjects moved from S to A, and the robot from R to A.

3. At position A, the subjects asked the question, and the robot answered. The robot could make its utterance with synthesized speech sounds. The content of the utterance was "Go forward, turn right, turn left, turn right, turn left, and then you'll be at the destination." The robot could make gestures while uttering this. In these experiments, we prepared six conditions under which the robot's gesture was changed.

4. The subjects tried to go to the lobby after receiving the robot's directions.

5. The experiments finished whether the subjects arrived at the lobby or gave up after losing their way. The subjects psychologically evaluated the robot through a questionnaire after the experiments finished.

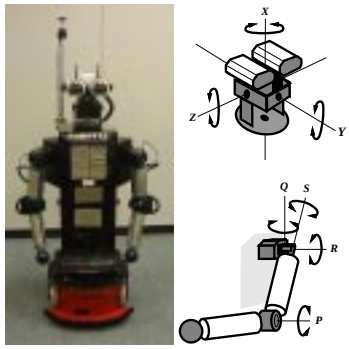

Figure 3: Outline of robot called "Robovie" (left), and robot's head and arm motion mechanisms (right).

Conditions We prepared the following six conditions from C-1 to C-6, which differed in terms of the robot's body movements (see Figure 5). The content of the utterance was the same under every condition.

C-1 (No gesture): The robot did not move.

C-2 (Absolute gesture): The robot raised its left arm leftward when telling the subject that he/she should go right, while it raised its right arm rightward when telling the subject that he/she should go left.

C-3 (Absolute gesture with gaze): In addition to C-2, the robot turned its eyes to the subject while making the utterance.

C-4 (Only aligned torso): The robot rotated so that it aligned its torso with the subject.

C-5 (Aligned gesture): In addition to C-4, the robot raised an arm forward telling the subject when he/she should go forward, rightward when the subject should go rightward, and leftward when the subject should go leftward.

C-6 (Aligned gesture with gaze): In addition to C-5, the robot turned its eyes to the subject while making the utterance.

Evaluations The results of the experiments were evaluated from the record of the subjects' behaviors and the answers of the questionnaire. In the questionnaire, the subjects were asked whether they understood the robot's utterance and to give a psychological evaluation of the robot on a seven-point scale for six items: *Familiarity*, *Sincerity*, *Reliability*, *Intelligence*, *Extroversion*, and *Kindness*.

Predictions

In the experiments, we gave evidence for the following three predictions derived from the proposed model. The more the robot's gestures increase systematically rather than randomly, i.e., the more the conditions shift in order from C-1 to C-6,

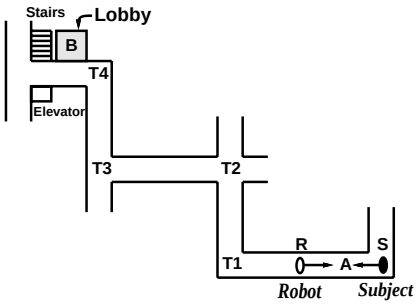

Figure 4: Experimental setup: arrangement of subject, robot, destination, and turns.

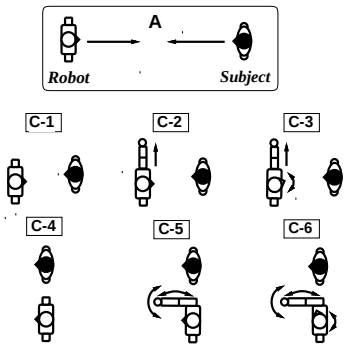

Figure 5: Outline of experimental conditions under changing robot gestures.

Prediction 1: the more the subjects' gestures will increase by entrainment and synchronization with the robot's, and consequently, a relationship will emerge from the mutual gestures.

Prediction 2: the easier the joint viewpoint will be obtained by the relationship.

Prediction 3: the easier the subjects will understand the utterance of the robot and arrive at the destination by using the obtained viewpoint.

Here, Predictions 1, 2, and 3 correspond to Formulae (3), (2), and (1) in the model of embodied communications, respectively.

Results

We give evidence for the three predictions in the order of Predictions 1, 3, and 2 to make the point of our argument clearer.

Verification of Prediction 1 From the observation results on the subjects' behaviors, we analyzed the subjects' gestures. First, the gestural arrangements that the subjects took were as we had expected (see Figure 5). Next, Figure 6 shows the ratio of appearances of the subjects' body movements under each condition. In this analysis, we classified the subjects into three categories: subjects who did not practice body movements at all (*Nothing*), subjects who only moved their hands (*Hand*), and subjects who moved and raised their hands

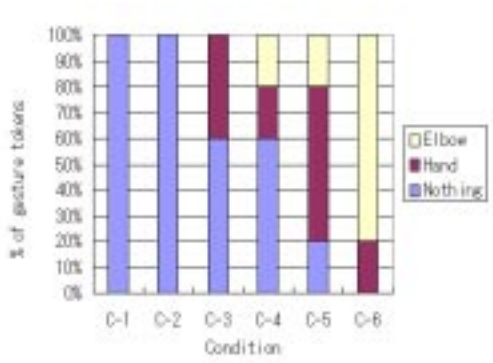

Figure 6: Results of subjects' body movements in human-robot interaction.

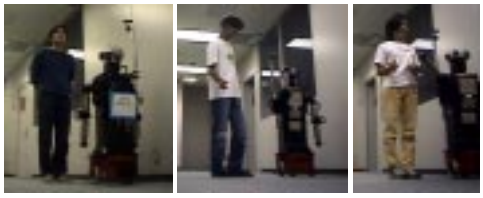

Figure 7: Photos of a subject under Condition C-1 (left) and two subjects under C-6 (center and right).

up to the elbow level (*Elbow*). In the analysis, a significant difference was found between the ratio of appearances of the subjects' body movements and the conditions ($\chi^2 = 25.210,\ p < .01$). In other words, the more the conditions shifted from 1 to 6 (i.e., the more the robot's gestures increased systematically), the more the subjects' gestures increased in sync. Moreover, the average scores for the numbers of times the subjects turned their eyes to the robot were higher when the robot turned to meet the eyes of the subjects (C-3 and C6).

We show appearances of the experiments in Figure 7. First, the left-hand side of Figure 7 shows the appearance of a subject not making any body movement and not turning his eyes to the robot at all (C-1). In contrast, the center of Figure 7 shows the appearance of a subject making an entrained body movement and turning his eyes to the robot (C-6). The right-hand side of Figure 7 also shows the appearance of a subject making an entrained body movement and turning her eyes in the same direction as the robot (C-6).

As a result of the observations described above, we could confirm that relationships emerged between the subjects and the robot from mutually entrained gestures. Consequently, Prediction 1 was supported.

Verification of Prediction 3 We recorded the time the subjects spent moving from A to B in Figure 4. Table 2 shows the average time and the number of subjects not arriving at B under each condition. Regarding the average time, no significant difference was found between the conditions. However, the average time in C-6 was the shortest.

Table 2: Average time until subjects' arrival at destination, and number of subjects not arriving at destination.

	C-1	C-2	C-3	C-4	C-5	C-6
Time to destination	69.5	71.3	67.7	70.2	66.8	65.4
Number of subjects not arriving	1	2	2	0	0	0

A noteworthy point is that a considerable number of subjects did not arrive at the goal in C-1, C-2, and C-3. The results of the questionnaire clearly showed that the subjects who did not arrive were unable to correctly understand the robot's utterance. One of the comments often heard was that they could not understand whether the robot's utterance including "left" and "right" meant the robot's or the subjects'. In other words, the reason why the subjects did not understand the utterance was that they could not obtain a joint viewpoint with the robot.

Consequently, the subjects who did manage to obtain a joint viewpoint could understand the robot's utterance and arrive at the goal, whereas the subjects who did not were unable to understand and arrive at the goal. Therefore, Prediction 3 was supported.

Verification of Prediction 2 First, we discuss obtaining a joint viewpoint from the aspect of body arrangement. Under the verification of Prediction 3, it was clear that all of the subjects could obtain a joint viewpoint when the robot aligned its body arrangement with the subject's to the destination (C-4, C-5, and C-6). In contrast, approximately one-third of the subjects could not obtain a joint viewpoint when the robot did not align its body arrangement (C-1, C-2, and C-3). Consequently, it is hard for subjects to obtain a joint viewpoint when no relationship emerges from the use of body arrangement.

Next, we discuss obtaining a joint viewpoint from the aspect of mutually entrained gestures such as synchronized arm movements and eye contact. As discussed in the verification of Prediction 1, from the results of the observed data, the more the robot's gestures increased systematically, the more the subjects' gestures did so. Moreover, from the results of the questionnaire, the more the conditions shifted from C-1 to C-6, the higher the average scores became (see Figure 8). In other words, the more the conditions shifted, the smoother the communications became. Based on this consideration, the relationship that emerged from the entrained gesture made it easier to obtain the joint viewpoint.

As a result of the above observations, Prediction 2 was supported. Consequently, the validity of our proposed model was given evidence by the three supported predictions.

Discussion and Conclusions

In this paper, we proposed a model of embodied communications focusing on body movements. Moreover, we explored the validity of the model through experiments on human-human communications involving giving/receiving route directions. The results of the experiments roughly supported our research direction. However, we could not investigate the details of the model because we were unable to manipulate the parameters in the experiments. Therefore, we carried out similar experiments using our implemented robot system. From the results of these experiments, we could give evidence for the validity of the model more appropriately.

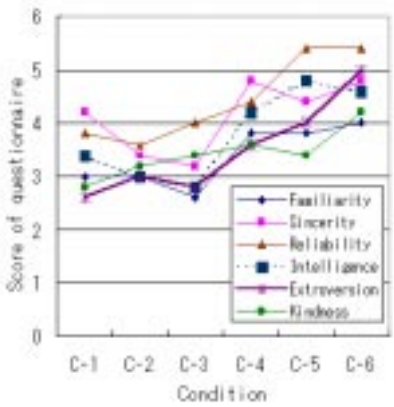

Figure 8: Results of subjects' psychological evaluations to a robot.

The contributions of our research should be viewed from two perspectives. First, our model of embodied communications suggests a new direction in research on communications. The target of previous research had mainly been the mechanism of verbal communications based on informatics approaches, e.g., Shannon's model. After that, McNeil's school pointed out that gestures are synchronized with speech. However, they have not yet modeled a whole conception of interactive communications that includes the function of embodiment. Our model gives a clue toward better understanding of such communications.

Moreover, the results of this research can be applied to interactive technologies between humans and artifacts. In other words, artifacts that can draw out human physical movements can make humans feel familiar with them. These cognitive engineering technologies enable us to develop an interface system and a robot system in the work's next generation.

References

Kita, S. (2000). Interplay of gaze, hand, torso orientation and language in pointing. In *Pointing: Where language, culture, and cognition meet.* Cambridge University Press, Cambridge.

McNeill, D. (1987). *Psycholinguistics: A new approach.* Harper & Row.

Ono, T., Imai, M., Ishiguro, H., & Nakatsu, R. (2001). Embodied communication emergent from mutual physical expression between humans and robots. *Transactions of Information Processing Society of Japan*, (submitted).