Timing Control of Utterance and Body Motion in Human-Robot Interaction

Kensaku Namera¹, Shoji Takasugi¹, Koji Takano¹, Tomohito Yamamoto² and Yoshihiro Miyake¹

Abstract— Towards developing robots that able to communicate 'naturally' with humans, we believe that it is important to study, and note any important interrelation between, both verbal and non-verbal information in human communication. The timing control model was developed based on our previous study that analyzed the utterance and body motions in the context of human-human communication. This paper presents the realization of this timing control model for the purpose of human-robot interaction, implemented on the Wakamaru robot platform and verifying the effectiveness of the model for it.

I. INTRODUCTION

Robots able to communicate with humans have been recently studied along with other remarkable technological developments. We think that “communication robots”, i.e. robots (especially humanoid) that their primary function is to share information and “naturally” interact with humans, will be able to assume the roles of partners with humans in the future. However, natural and smooth communication between humans and robots has not been achieved yet. So, the main purpose of this study is toward realizing a more natural and smooth interaction between them.

Therefore, it is necessary to first clarify the mechanism of human communication. In human conversation, both verbal and non-verbal information play an important role [1]. This non-verbal information includes not only visual processing on the gestures and the gaze, but also the prosodic information of voice, especially the temporal structure of speech such as the utterance timing which has been considered as an important element for smooth communication.

For instance, Condon et al. [2] showed that the interaction between speech rhythm and body rhythm is the key role in the communications between mother and infant. Watanabe et al. [3] reported that entrainment is observed in the rhythm of utterance and nod, and applied it to many robot interfaces

[4]. Matarazzo et al. [5-6] clarified that the duration of utterance, the speed of utterance, and the reaction time are tuned among speakers. Nagaoka et al. [7, 8] reported the synchronization phenomenon between switching pauses and utterance speed. Breazeal et al. [9] showed how people read and interpret non-verbal cues from a robot in order to coordinate their behavior that improved teamwork efficiency and robustness. Dautenhahn et al. [10] described that imitation was important for social learning.

However, verbal and non-verbal information in human communications have been studied independently, with any interrelations between them not clarified in previous research. Yamamoto et al. [11, 12] have analyzed the temporal structure between utterance as verbal information and body motion as non-verbal information in human dialogue, and a timing control model of conversation was proposed. Based on that report, we aim to realize a timing control model of the utterance and body motion for more natural and smooth responses in human-robot interaction, and to implement this model on a physical robot system in order to realize and estimate its effectiveness.

We explain the analysis of human-human communication and its modeling in section II, describe the results of the model implementation for human-robot interaction and communication in section III, explain the experiment of evaluation of impression and describe the results of it in section IV, and finally summarize the work in section V.

II. TIMING CONTROL IN HUMAN COMMUNICATION

A. Method

In the experiment presented by Yamamoto et al. [11], the dialogue consisted of an instruction utterance and a response utterance. In each experiment session two persons (instructor and subject) sit across a desk, on which there are ten block objects (as shown in Fig.1), and repeat the following conversation (in Japanese) ten times, one for each object.

1. The instructor says "Would you please pick up that [block]."
2. The subject answers, "Yes" and picks up one block.

¹K. Namera, S. Takasugi, K. Takano and Y. Miyake are with the Dept. of Computational Intelligence & Systems Science, Interdisciplinary Graduate School of Science & Engineering, Tokyo Institute of Technology, Japan. (corresponding author email: miyake@dis.titech.ac.jp)
²T. Yamamoto is with the Dept. of Information Engineering and Computer Science, Graduate School of Engineering, Kanazawa Institute of Technology, Japan.
In order to clarify the influences of the change in the utterance speed on the temporal structure of the dialogue, the utterance speed of the instructor was intentionally changed in this experiment. To analyze the temporal structure of this dialogue, the following four behaviors were used as indices of timing control. Instructor was characterized by utterance, and instructed subject was characterized by utterance, nodding and grasping.

**B. Results**

The timing chart of the dialogue between instructor and subject is shown in Fig. 2. In this figure, the time of speech duration of the instructor is indicated by $S_d$, and the “switching pause” which is defined as the time interval between the end of instruction utterance and the start of response utterance is shown as $P_d$. The time difference between the start of nodding and the start of the response utterance (“nodding pause”) is defined as $N_p$. The time difference between the start of grasping and the start of the response utterance (“grasping pause”) is defined as $G_p$. 

**Fig. 3.** Correlation between the duration of instruction utterance ($S_d$, horizontal axis) and the duration of the switching pause ($P_d$, vertical axis) in Yamamoto et al. [11, 12].

**Fig. 4.** Correlation between the duration of the switching pause ($P_d$, horizontal axis) and the duration of the nodding pause ($N_p$, vertical axis) in Yamamoto et al. [11].

**Fig. 5.** Correlation between the duration of the switching pause ($P_d$, horizontal axis) and the duration of the grasping pause ($G_p$, vertical axis) in Yamamoto et al. [11].
Three correlations were found. The first one is the relationship between $S_d$ and $P_d$, the second one is the relation between $P_d$ and $N_p$, and the last one is the relation between $P_d$ and $G_p$. Figure 3 shows the positive correlation between $S_d$ and $P_d$, indicating that when the duration of an instruction utterance becomes longer, the switching pause also becomes longer. Figure 4 shows the positive correlation between $P_d$ and $N_p$, indicating that when the duration of the switching pause becomes longer, the duration of the nodding pause also becomes longer. Figure 5 shows the positive correlation between $P_d$ and $G_p$, indicating that when the duration of the switching pause becomes longer, the duration of the grasping pause also becomes longer.

C. Timing Control Model

From these correlations, Yamamoto et al. [11, 12] proposed the following timing control model:

\[ P_d = a*S_d + b, \]  
\[ N_p = c*P_d + d, \]  
\[ G_p = e*P_d + f, \]

where $a$, $b$, $c$, $d$, $e$ and $f$ are Real numbers, and $a$, $c$, $e > 0$. In the next section, we present the implementation of such a timing control model for the purposes of human-robot interaction and communication in a robotic system.

III. IMPLEMENTATION FOR HUMAN-ROBOT INTERACTION

A. Wakamaru Robot

We implemented this timing control model on the communication robot “Wakamaru” (developed by the Mitsubishi Heavy Industries Inc.), shown in Fig. 6. Wakamaru has a multiprocessor CPU composition, and uses the Linux OS. It can recognize spoken words by a wordspotting method and can in turn speak by reading text data. It has a neck with 3 DOF (degrees of freedom), 2 arms with 4 DOF each and wheels with 2 DOF. When the implemented program that included the timing model was executed, all other previously equipped programs (that usually run in parallel) were shut down.

B. Implementation of the timing control model

We designed a modified dialogue task to implement the model. A human subject sits across a table (which this time has only one object on it) from Wakamaru, and makes the following conversation:

1. The subject instructs Wakamaru “Would you pick up that [object].”
2. Wakamaru then replies, “Yes” and makes a pointing gesture towards the object.

For technical reasons, we could not have the robot actually grasp and pick up the object.

There were two additional issues with the implementation. One was regarding the motion control mechanism that Wakamaru uses. We could not separate the execution of the nodding and pointing gestures (i.e. define the $G_p$ timing independent of $N_p$). So in our program, the nodding and pointing are done at the same time. Therefore equation (3) needs to be modified as follows:

\[ G_p = N_p. \]

The other issue was concerning the voice recognition. We needed to obtain the precise value of duration time of the instruction utterance. In this study, we calculated it from the time difference between the start and the stop timings of voice recognition module. But, the sentence “Would you please pick up that [object]” was not easily recognized by the voice recognition system (VORERO, Asahikasei Inc.) equipped on Wakamaru. So we modified the sentence as “Would you pick up that [object]” to be recognized more easily by the system.

C. Calibration of Implemented System

When implementing the timing control model in the Wakamaru robot, we did a series of sessions in which we measured the difference between the target values (as measured by an observer during the session or estimated from the model) and the actual values (as measured by the Wakamaru system) of timings, in order to calibrate the model.

1) Duration of subject utterance: $S_d$

First, we calibrated the duration of subject’s utterance measured by Wakamaru. Figure 7 shows the relationship between the actual duration of utterance by the human subject and the measured duration of the same utterance by Wakamaru (with a superimposed regression line). As shown

5 The phrase is “Sore, tottekudasai” in Japanese.

4 The phrase is “Sore, tyohdai” in Japanese.
in this figure, the error of the measured utterance duration was very small; the regression line is near the “y=x” line, indicating that the error is sufficiently small.

2) Duration of switching pause: \( Pd \)

Next, we calibrated the duration of the switching pause as realized by Wakamaru. Figure 8 shows the relationship between the switching pause duration in the timing control model and the actual duration of the switching pause by Wakamaru. As shown in this figure, the error was about 700ms (distance of the superimposed regression line from the “y=x” line). This means that the timing control system in the Wakamaru robot has a time delay of about 700ms.

3) Duration of nodding pause: \( Np \)

Finally, we calibrated the time difference between the start of the nodding gesture and the start of the utterance, both by Wakamaru. Figure 9 shows the relationship between the time difference from the timing control model and the observed time difference by Wakamaru. As shown in this figure, the error of the observed time difference was very small (distance of the superimposed regression line and the “y=x” line), indicating that the control of the timing system is operating correctly.

4) Compensation of timing lag

From the above calibrations, it was clarified that the time lag of the timing control system equipped in Wakamaru should be considered, and especially the time lag in the duration of the switching pause \( Pd \). So, we compensate by using the following parameter values for the timing control model as implemented in the Wakamaru robot:

\[
Pd = 0.9677 \times Sd - 483.87, \quad (5) \\
Np = 0.6667 \times Pd, \quad (6) \\
Gp = Np. \quad (7)
\]

![Fig. 7. The error between the duration of utterance (\( Sd \)) as estimated by the model (horizontal axis) and the actual observed value measured by the Wakamaru robot (vertical axis). The superimposed regression line (solid) is very close to the \( y=x \) line (dotted), indicating only a small error.](image)

![Fig. 8. The error between the duration of the switching pause (\( Pd \)) as estimated by the model (horizontal axis) and the actual value measured from the Wakamaru robot (vertical axis). The distance of the superimposed regression line (solid) from the \( y=x \) line (dotted), indicate a lag of about 700ms.](image)

![Fig. 9. The error between the nodding pause (\( Np \)) as estimated by the model (horizontal axis) and the actual value measured from the Wakamaru robot (vertical axis). The superimposed regression line (solid) is very close to the \( y=x \) line (dotted), indicating only a small error.](image)

IV. EXPERIMENT

We conducted an experiment to verify the effectiveness of the timing control model for realizing natural and smooth interaction between human and robot.

A. Subjects

Since it is known that the time perception changes by aging [13], we selected two generation groups as the subjects, elderly persons and young persons. For elderly persons, 18 male & female subjects (age 66-74, average age 69) were used. For young persons, 18 male & female university students (average age 23) were used.
B. Experimental Setup

The experimental setup is shown in Fig. 10. The response actions of Wakamaru are shown in Fig. 11. The task was as described in section III.B, and $Pd$ and $Np$ were controlled in this experiment. We used a soundproof room to prevent surrounding noise.

In the experiment, we compared “Fix” ($Pd$ and $Np$ are fixed) and “Model” ($Pd$ and $Np$ are controlled based on the model described in the previous chapter) conditions. The parameters for the “Model” condition were the same as in section III.B but the actual system was as follows, because of the time lag that was also described in section III.C.

$$Pd = 0.9347*Sd + 110.15,$$
$$Np = 0.6667*Pd.$$

The parameters for the “Fix” condition were as follows,

$$Pd = 1000,$$
$$Np = 800.$$

Subjects answered to the questionnaires shown in Fig. 12, after this experiment. The first 3 questions were about their impression of the experiment, and the remaining questions were about the psychological time and space.

C. Experimental Procedure

Subjects conducted the task under the “Model” and “Fix” conditions for six times respectively. We did not inform the subjects about the two different conditions. The order was random for each subject. Subjects were instructed to employ three different speeds of when speaking, “fast”, “usual” and “slow”. They were then asked to evaluate their impression of the response timing of the robot, immediately after the end of each task.

![Fig. 10. Experimental system](image1)

![Fig. 11. Response actions of Wakamaru](image2)

![Fig. 12. Questionnaires](image3)

![Fig. 13. Questionnaire results (elderly person)](image4)

![Fig. 14. Questionnaire results (young person)](image5)
D. Results and Discussion

The results for the elderly person group are shown in Fig. 13. The effectiveness of “Model” was remarkably high in all questions except the fifth. Therefore, it was shown that “Model” was effective for the elderly person group. The results for the young person group are shown in Fig. 14. The effectiveness of “Fix” was higher than that of “Model” in all questions, though the difference was small. Therefore, we could say that the effectiveness of “Model” was not seen by the young person group.

These results might reflect the different attitudes between elderly and younger people. Elderly people prefer someone that they spoke to slowly, to also reply slowly to them. In contrast, younger people prefer a fast response, regardless their own speed of speech. Therefore the effectiveness of the “Model” condition was not seen by the young person group, but only by the elderly person group. But, if we can delete the lag and Wakamaru can response faster, the results for young person might be changed.

V. CONCLUSION

In this study, we focused on the relationship between verbal and non-verbal information. We realized a timing control model of the utterance and the body motion implemented on the Wakamaru robot for the purpose of human-robot interaction and communication. The model was proposed by Yamamoto et al. [11, 12] for achieving a more natural communication based on human-human conversation. We verified that it is effective between human-robot interaction as well. We expect the development of robot technology that can realize a smooth and natural communication between human and robot to become possible by investigating the interaction in human communication.

ACKNOWLEDGEMENTS

This work was funded by Project for Strategic Development of Advanced Robotics Elemental Technologies of NEDO (New Energy and Industrial Technology Development Organization), Japan. And the authors would like to thank Mitsubishi Heavy Industries, Ltd. for providing the Wakamaru robot.

REFERENCES