Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, Technische Universität München, Munich, Germany, August 1-3, 2008

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

Kensaku Namera[1], Shoji Takasugi[1], Koji Takano[1], Tomohito Yamamoto[2] and Yoshihiro Miyake[1]

Abstract— **Towards developing robots that are capable of communicating in 'natural' ways 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 model of a timing control system was developed based on a 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 system for the purpose of human-robot interaction, implemented on the Wakamaru robot platform.**

I. Introduction

Robots able to communicate with humans have been recently studied along with other remarkable technological developments. In these years, humanoid robots which show human-like behavior [1], robots that can support humans in their everyday-life activities or can take the role of an "entertainment robot" [2, 3] and also "toy robots" which can recognize the faces and voices of children when playing with them [4] have already been proposed. We think that "communication robots", 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 purpose of this study is to realize more natural and smooth communication in such interaction between human and robots.

Therefore, it is necessary to first clarify the mechanism of human communication. In our human conversation, both verbal information and non-verbal information plays an important role [5]. 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.* [6] showed that the interaction between the speech rhythm and the body rhythm was the key role in the communications between mother and child. Watanabe *et al.* [7] reported that entrainment is observed in the rhythm of the utterance and the nod, and applied it to many human interfaces [7]. Matarazzo *et al.* [8-10] clarified that the duration of utterance, the speed of utterance, and the reaction time are tuned among speakers. Nagaoka *et al.* [11, 12] reported the synchronization phenomenon between switching pauses and utterance speed.

However, verbal and non-verbal information in human communications have been studied mostly independently, with any interrelations between them not clarified in previous research.

Yamamoto *et al.* [13, 14] 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 human conversation was proposed. Based on that report, we aim to construct a timing control model of the utterance and body motion for realizing a more natural and smooth response in human-robot interaction, and to implement the model on a real robot system to 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, and finally summarize the work in section IV.

II. Timing Control in Human Communication

A. Methodology

In the experiment presented in Yamamoto *et al.* [13], the dialogue consisted of an instruction utterance and a response utterance used in the experiment. In each experiment session two persons (the instructor, which was one of the researchers, and a 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.

[1] K. 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)

[2] T. Yamamoto is with the Dept. of Information Engineering and Computer Science, Graduate School of Engineering, Kanazawa Institute of Technology, Japan

978-1-4244-2213-5/08/$25.00 ©2008 IEEE

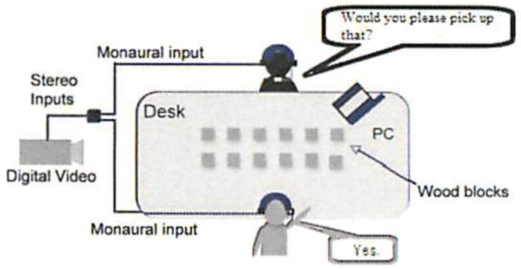

Fig. 1. The experimental setup used in Yamamoto *et al.* [13, 14]. See section II.A. for details.

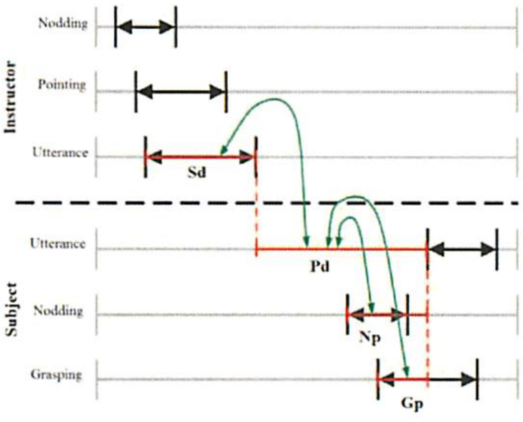

Fig. 2. The timing chart of the dialogue between instructor and subject used in Yamamoto *et al.* [13]. See section II.B. for details.

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 by them in this experiment.

To analyze the temporal structure of this dialogue, the following six behaviors were used as indices of timing control. Instructor subject was characterized by *utterance*, *nodding* and *pointing*, 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 *Sd*, 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 *Pd*. The time difference between the start of nodding and the start of the response utterance ("nodding pause") is defined as *Np*. The time difference between the start of grasping and the start of the response utterance ("grasping pause") is defined as *Gp*.

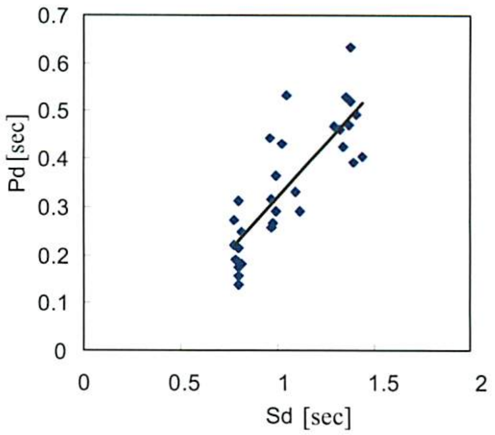

Fig. 3. Correlation between the duration of instruction utterance (*Sd*, horizontal axis) and the duration of the switching pause (*Pd*, vertical axis) in Yamamoto *et al.* [13, 14].

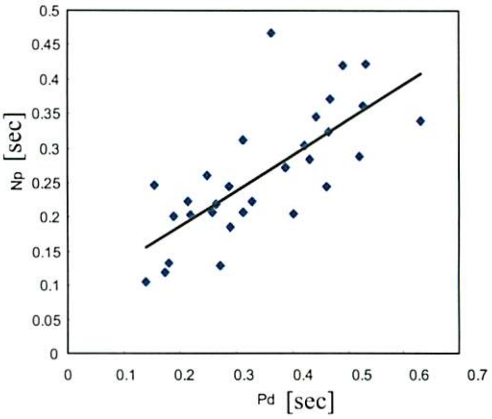

Fig. 4. Correlation between the duration of the switching pause (*Pd*, horizontal axis) and the duration of the nodding pause (*Np*, vertical axis) in Yamamoto *et al.* [13].

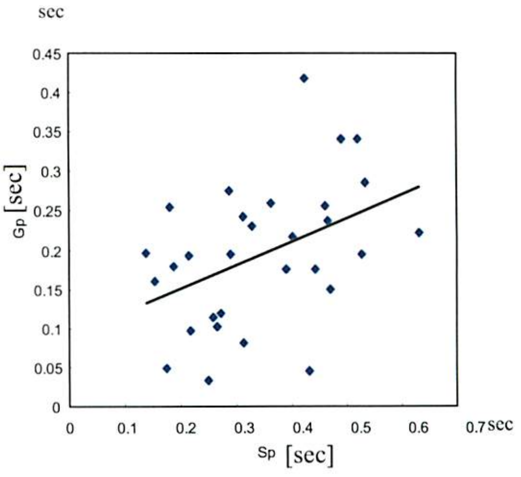

Fig. 5. Correlation between the duration of the switching pause (*Pd*, horizontal axis) and the duration of the grasping pause (*Gp*, vertical axis) in Yamamoto *et al.* [13].

Three correlations were found. The first one is the relationship between *Sd* and *Pd*, the second one is the relation between *Pd* and *Np*, and the last one is the relation between *Pd* and *Gp*.

Fig. 3 shows the positive correlation between *Sd* and *Pd*, indicating that when the duration of an instruction utterance becomes longer, the switching pause also becomes longer. Fig. 4 shows the positive correlation between *Pd* and *Np*, 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 *Pd* and *Gp*, 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.* [13, 14] proposed the following timing control model:

$$Pd = a*Sd + b \qquad (1)$$
$$Np = c*Pd + d \qquad (2)$$
$$Gp = e*Pd + f \qquad (3)$$

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.) as 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 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.

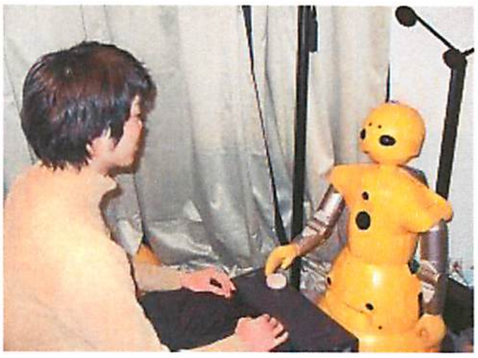

Fig. 6. Wakamaru robot in human robot interaction.

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

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

$$Gp = Np \qquad (4)$$

The other problem 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 the voice recognition module. But, the sentence "Would you please pick up that [object]"[3] 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]"[4] 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: Sd

First, we calibrated the duration of subject utterance measured by Wakamaru. Fig. 7 shows the relationship between the actual duration of utterance by the human

[3] The phrase is "Sore, tottekudasai" in Japanese.

[4] The phrase is "Sore, tyohdai" in Japanese.

subject and the measured duration of that utterance by Wakamaru (with a superimposed regression line). As shown in this figure, the error of the measured utterance duration was very small; the regression line is near the "y=x" line, which indicates that the error is very small.

2) Duration of switching pause: Pd

Next, we calibrated the duration of the switching pause as realized by Wakamaru. Fig. 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. Fig. 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

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:

$$Pd = 0.9677*Sd - 483.87 \qquad (5)$$

$$Np = 0.6667*Pd \qquad (6)$$

$$Gp = Np \qquad (7)$$

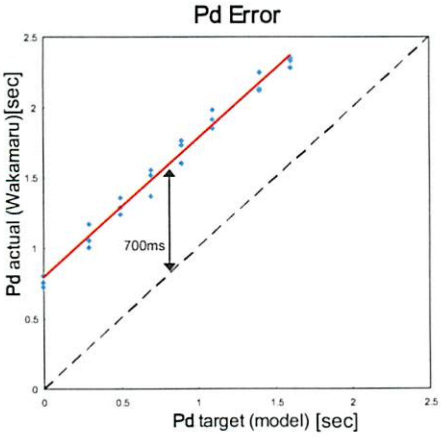

Fig. 7. The error between the duration of utterance (*Sd*) as estimated by the model (horizontal axis) and the actual observed value measured by the Wakamaru robot (vertical axis). The superimposed regression line (solid) is very close to the *y=x* line (dotted), indicating only a small error.

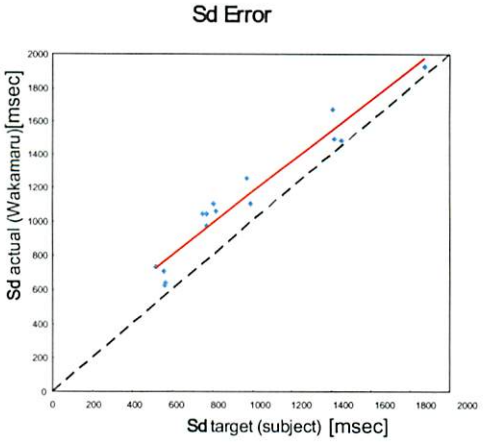

Fig. 8. The error between the duration of the switching pause (*Pd*) as estimated by the model (horizontal axis) and the actual value measured from the Wakamaru robot (vertical axis). The distance of the superimposed regression line (solid) from the *y=x* line (dotted), indicate a lag of about 700ms.

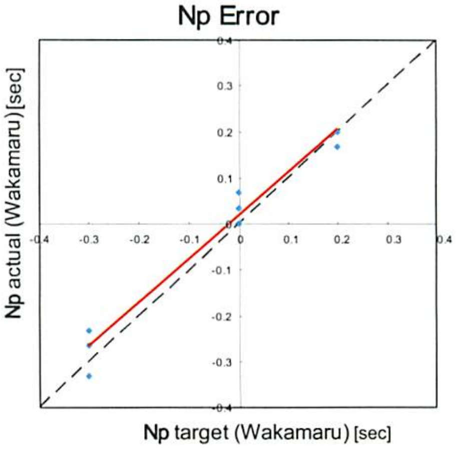

Fig. 9. The error between the nodding pause (*Np*) as estimated by the model (horizontal axis) and the actual value measured from the Wakamaru robot (vertical axis). The superimposed regression line (solid) is very close to the *y=x* line (dotted), indicating only a small error.

Recently, we are estimating the effectiveness of this timing control model in the interaction between Wakamaru robot and human, and the results will be clarified till the conference of ROMAN2008.

IV. Conclusions

In this study, we focus on the relationship between verbal and non-verbal information, by realizing 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.* [13, 14] for realizing a more natural communication based on human-human conversation.

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.

REFERENCES

[1] Sakagami,Y., Watanabe, R., Aoyama, C., Matsunaga, S., Higaki, N and Fujimura, K. : The intelligent ASIMO; System overview and integration, Int. Conf. on Intelligent Robots and Systems (IROS'02), pp. 2478-2483, 2002.

[2] Kanda, T., Ishiguro, H., Ono, T., Imai, M., Maeda, T. and Nakatsu, R. : Development of 'Robovic' as platform of everyday-robot research, The transactions of the Institute of Electronics, Information and Communication Engineers. D-I, Vol.J85-D-I, No.4, pp.380-389, 2002.

[3] Kawauchi, N., Koketsu, Y., Nagashima, T., Onishi, K. and Hiura, R. : Home-use robot "Wakamaru, Mitsubishi Heavy Industries Technical Review. Vol.40, No.5. 2003.

[4] Sato, M., Sugiyama, A., Houshuyama, O., Yamashita, N., Ohnaka, S. and Fujita,Y: The voice interface of personal robot, PaPeRo, The Journal of the Acoustical Society of Japan, Vol.62, No.3, pp.173-181, 2006.

[5] Daibou, I. : "Shigusano komyunikeishon – hito ha shitashimi wo doutsutaeauka - " in Japanese , Saiensu-sha, 1998.

[6] Condon, W.S. and Sander, L.W.: Neonate movement is synchronized with adult speech, Science, Vol.83, pp.99-101, 1974.

[7] Watanabe, T., Okubo, M., Nakashige, M. and Danbara, R. : An embodied interaction system based on speech by using inter-actor, The Transaction of Human Interface Society, Vol.2, No.2, pp.21-29, 2000.

[8] Matarazzo, J.D. and Wiens, A.N.: Interviewer influence on durations of interviewee silence, Journal of Experimental Research in Personality, Vol.2, pp.56-69, 1967.

[9] Matarazzo, J.D., Weitman, M., Saslow, G. and Wiens, A.N. : Interviewer influence on durations of interviewee speech, Journal of Verbal Learning and Verbal Behavior, Vol.1, pp.451-458, 1963.

[10] Matarazzo, J.D. and Wiens, A.N. : Interviewer influence on durations of interviewee silence, Journal of Experimental Research in Personality, Vol.2, pp.56-69, 1967.

[11] Nagaoka,C. : Taijin komyunikeishon niokeru higenngokoudouno 2shakann sougoeikyou in Japanese, Taijinsinrigakukenkyu, Vol.6, pp. 101-112, 2006.

[12] Nagaoka, C., Komori, M., Nakamura, T. and Draguna, M.R. : Effects of receptive listening on the congruence of speakers' response latencies in dialogues, Psychological Reports, vol.97, pp.265-274, 2005.

[13] Yamamoto,T *et al.* : Analysis of utterance and body motion timing (title tentative) : Technical report in Kanazawa Institute of Technology, in preparation.

[14] Yamamoto, T., Hirano, T., Kobayashi, Y., Takano, K., Muto, Y. and Miyake, Y. : Two types of correlation of utterance timing in dialogue, Human Interface symposium 2007.