Robot Motion Control using Listener's Back-Channels and Head Gesture Information

Tsuyoshi Tasaki, Takeshi Yamaguchi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan
{tasaki, takeshi, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp

Abstract

A novel method is described for robot gestures and utterances during a dialogue based on the listener's understanding and interest, which are recognized from back-channels and head gestures. "Back-channels" are defined as sounds like 'uh-huh' uttered by a listener during a dialogue, and "head gestures" are defined as nod and tilt motions of the listener's head. The back-channels are recognized using sound features such as power and fundamental frequency. The head gestures are recognized using the movement of the skin-color area and the optical flow data. Based on the estimated understanding and interest of the listener, the speed and size of robot motions are changed. This method was implemented in a humanoid robot called SIG2. Experiments with six participants demonstrated that the proposed method enabled the robot to increase the listener's level of interest against the dialogue.

1. Introduction

The growing demand for family robots like SONY AIBO which interests human at home needs to communicate with people smoothly. Since it is difficult for robots to completely understand what user says, much attention has been paid to "paralanguage", which supports smooth information transfer in dialogue [1]. Paralanguage is classified into audio expressions and visual expressions. For example, audio expressions include back-channels (prosody information in utterances), and visual expressions include head gestures. "Back-channels" are defined as sounds like 'uh-huh' uttered by a listener during a dialogue, and "head gestures" are defined as nod and tilt motions of the listener's head.

Previous studies have demonstrated only the control of dialogue content based on the content of listener's speech and context [2][3]. There have been few reports using the robot systems, which can control not only the utterance speed and utterance volume but also the body motions. We propose a method that enables a robot to change its utterance speed, volume and its body motion based on the listener's understanding and interest estimated by listener's back-channels and head gestures.

Our aim is to enable a robot keep the listener's interest during the interaction by dynamical adjustment of the body motions controlled based on the estimation of listener's condition.

Section 2 describes our method for recognizing back-channels and head gesture and describes our design for controlling the robot's motion based on the listener's understanding and interest. Section 3 describes the results of our experiments and evaluates the proposed method. Section 4 concludes this paper and mentions future work.

2. Estimation of Listener's Understanding and Interest and Robot Motion Control

This section describes our method for estimating the listener's understanding and interest and controlling the robot's motions. In this paper, "listener's understanding and interest" are defined as combination of the degree of understanding and the listener's level of interest. Section 2.2 describes it in detail.

Figure 1 shows our system. There are two main inputs in the system. One is sound detected by microphone in back-channels recognition client. The other is image detected by camera in head gestures recognition client. The system estimates listener's understanding and interest using these information and memorizes listener's action timing in server. Finally, it controls robot's motion in motor client and utterance client.

2.1. Back-Channels and Head Gestures

We use back-channels and head gestures to estimate the listener's understanding and interest.

Maynard [4] stated that Japanese back-channels have six functions with vagueness. We identified three elements as functions of the back-channels by summarizing the functions proposed by Maynard.

- C: Continuance Demand (Demand for Speech Continuance)
- E: Emotional Expression
- D: Demand for Information Addition or Correction

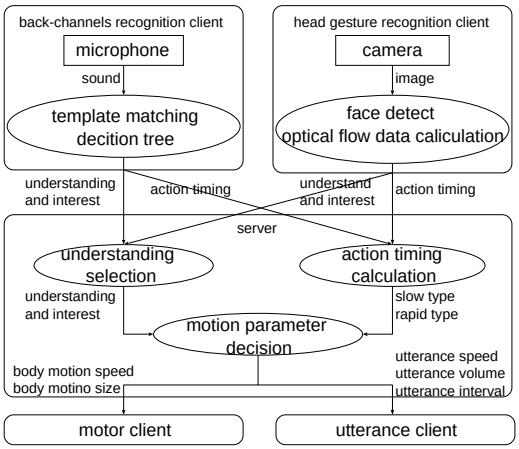

Figure 1: System Structure

As the head gestures, we use a nod motion (vertical movement) and a tilt motion (horizontal movement). In a dialogue, a vertical movement shows "affirmation" and a horizontal movement shows "negation".

2.2. Listener's Understanding and Interest

Among three functions of back-channels, C and D cannot appear at the same time, while E can appear at the same time with C or D. Moreover, affirmative and negative head gestures cannot appear at the same time. They are supplement functions of C and D.

Therefore, two factors can be extracted from the three functions of back-channels and two head postures. One shows how well the listener understands the dialogue (understanding degree), and the other shows how much the listener is interested in the dialogue (interest degree). The understanding degree has two states: (1) "understand" and (2) "don't understand". The interest degree has three states: (1) friendly and positive where the listener is interested in the dialogue; (2) boring and negative condition where the listener is not interested in the dialogue; (3) normal condition which is neither positive nor negative.

In this study, the listener's understanding and interest are categorized into six states by combining understanding degree (H: high-level understanding, L: low-level understanding) with interest degree (I: Interested, N: Normal, B: Bored). For example, condition (H, B) means that the listener understands the dialogue, but is not interested in it.

Table 1: Correspondence between condition, back-channels, and head gesture

	understanding		interest	
	H	L	I	B
back-channels	C	D	E	E
head gesture	vertical	horizontal	-	-

Table 1 shows the correspondence between condition (understanding degree, interest degree), back-channels functions, and head movement. H corresponds to function C, and L corresponds to function D. Moreover, interest corresponds to function E. A vertical head movement corresponds to H and a horizontal head movement corresponds to L.

2.3. Estimation of Listener's Understanding and Interest

The recognition system of back-channels was developed using learning data of six dialogues which includes 376 back-channels (about 3 minutes total duration).

2.3.1. Back-channels labeling

At first, we labeled the back-channels in the learning data based on the six categories of understanding degree and interest degree.

Table 2 shows the results of labeling in each listener's conditions. The back-channels can be interpreted differently by different people. Therefore, two people (Labeler A and Labeler B) labeled these back-channels. The back-channels which both labeler corresponded were used for learning of template matching.

Table 2: Results of labeling

condition	Labeler A	Labeler B	corresponded
(H, I)	61	42	32
(H, N)	273	303	245
(H, B)	4	4	4
(L, I)	2	4	0
(L, N)	14	17	11
(L, B)	0	0	0
Not sure	22	6	-

The rate of corresponded labeling was 77.7%. The corresponded data for (L, I) was zero because people have trouble agreeing on this combination. The (L, B) combination did not arise because the dialogue is a cooperative dialogue between people. Therefore, these two combinations are neglected in this paper.

2.3.2. Back-channels and Head Gestures Recognition

The back-channels are recognized using template matching and a decision tree. Template matching uses two features: the utterance time and the utterance time beyond the initial standup power. The decision tree uses the results of the template matching and the three other features (the fundamental frequency form, the voiced section number, and the first voiced section inclination). Figure 2 shows the five features. The initial standup power is the power between the start of an utterance and first power falling. The voiced section is the section in which the fundamental frequency is obtained. The fundamental frequency form is the total approximate inclina-

tion of all voiced sections.

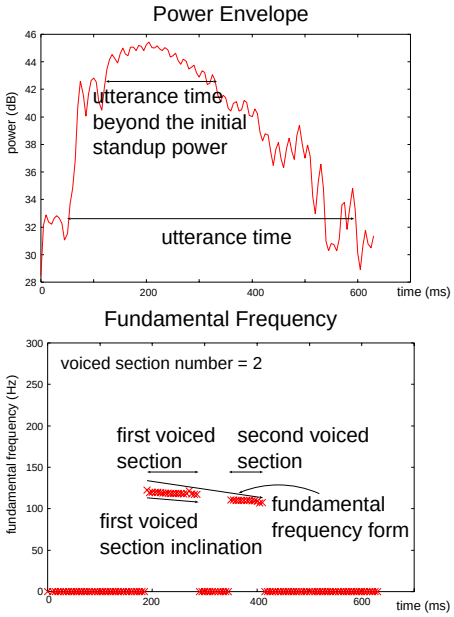

Figure 2: Five features for recognizing back-channels

Table 3 shows the recognition results for four cross-validation experiments. When the distance from the back-channels features to all the templates is very far, we regard the back-channels as 'noise'. In this paper, we recognize only four conditions, (H, I), (H, N), (H, B), and (L, N). (Refer to section 2.3.1.)

Table 3: Results of back-channels recognition

input ＼ output	(H, I)	(H, N)	(H, B)	(L, N)	noise	rate (%)
(H, I)	23	7	1	0	1	71.9
(H, N)	12	195	4	24	10	79.6
(H, B)	0	0	4	0	0	100
(L, N)	0	0	0	7	4	63.6

The head gestures are recognized based on optical flow data and the movement of the skin-color area, which is a standard method [1]. The rates for recognizing vertical and horizontal movements were both about 80%.

When the understanding degree obtained by back-channels recognition differed from that obtained by head gestures, the one including L was used as the result, because the condition that the listener does not understand should be avoided.

2.4. Robot Motion Control

This section describes how the motion of the robot is changed based on two kinds of information. One is the timing of the listener's actions and the other is listener's understanding and interest recognized by the proposed method.

First, the function based on listener's action timing is described.

By referring to [7], it is thought to be effective for robot to make actions after the listener's utterance so as to emphasize the utterance end.

Back-Channels and head gestures actions of human usually occur between 200 and 300 ms after an utterance is finished [5][6]. Therefore, if the action occurs more than 300 ms later, the robot should speak more slowly because the listener seems to want to speak slowly. Conversely, if it occurs less than 200 ms later, the robot should speak more rapidly because the listener seems to want to speak rapidly.

Table 4 shows parameters for the speed control based on listener's action timing.

Table 4: Change in speed parameter based on listener's action timing

action timing	less than 200 ms	more than 300 ms
speed	up	down

Next, the function based on listener's understanding and interest is described.

People automatically change utterance speed, utterance volume, and body motion based on the listener's understanding and interest. For example, the body motions or utterance volume are often larger and the utterance speed is reduced when the listener appears not to understand. Therefore, it is very important to change these parameters depending on whether the listener is interested or not and whether the listener understands the dialogue or not. The robot's movements are large and rapid if the listener looks interested, and they are smaller and slower if the listener looks uninterested.

In our method, the robot can change the speed parameter and the size parameter. The speed parameter affects the utterance speed, utterance interval, and body motion speed. The size parameter affects the utterance volume and body motion size. Table 5 shows control of both parameters based on the listener's understanding and interest. This is determined based on human's communication as mentioned above.

Table 5: Change in speed and size parameters based on listener's condition

condition	(H, I)	(H, N)	(H, B)	(L, N)
speed	up	-	down	down
size	up	-	down	up

By summarizing above two functions, the equation of robot control is described as follows:

$$\frac{dF}{dt} = c\Delta_t + c\Delta_{uF}, \frac{dA}{dt} = c\Delta_{uA}$$

F(t) is a speed parameter and A(t) is a size parameter. F(t) is changed by Δ_t which is variable based on listener's action timing, and Δ_{uF} which is variable based on listener's

understanding and interest. Δ_{uA} is variable to change A(t) based on listener's understanding and interest. ($10 \leq A(t) \leq 100, 10 \leq F(t) \leq 100, c : constant$)

3. Experiment and Result

We implemented our method in a humanoid robot, SIG2 (Figure 3). It has the soft skin with touch sensors covering whole body, and two microphones in external acoustic meatuses in side of the ears. Also, it has three degrees of freedom in the neck and one degree of freedom in the waist.

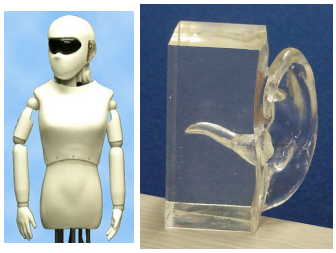

Figure 3: SIG2

We carried out experiments with six participants who were one female and five males.

The behaviors of each participant listening to the robot's speech were compared in two different cases. In first case, SIG2 changed its motion based only on the listener's action timing. In second case, SIG2 changed its motion based on both the listener's action timing and understanding and interest. The dialogue was the same for both situations and lasted about five minutes. Although the participants knew that SIG2 has a function for recognizing back-channels and head gestures, they did not know its detail functions and differences of two cases.

During the experiments, SIG2 used its own ears to recognize listener's back-channels and expressed its conditions by its utterance or nod and tilt motion of the head controlled by the proposed method described in Section 2.4. Listener's head gestures are recognized by the camera fixed outside of the robot.

Table 6 shows the average number of times of the listener's actions with interest, the distribution, and the rate for all participants. The number of action was counted by the experimenters.

Table 6: Comparing listener's interest

	average	distribution	rate(interested/all)
only timing	7.0	4.3	16%(42/265)
timing and condition	15	23	34%(92/269)

It is confirmed that people become more interested in a dialogue when the robot changes its motion based on both the listener's understanding and interest and the listener's action timing compared to when it changes its motion based only on the action timing. We also got almost same result from the questionnaires. This is because a listener gets interested more easily when the robot changes its motion based on the listener's understanding and interest.The distribution in the case based on action timing and condition is large because of individual difference.

When listener becomes interested in the dialogue with robots, he/she will listen to its speech. As a result, it can be said the robot can achieve smooth communication.

4. Conclusion

We proposed the method controlling the robot motion during dialogue based on user's backchannels and gestures. As the result of experiments, it was confirmed that people were more interested in the robot dialogue.

Future work includes design of more kinds of robot motions and enabling a robot to predict human actions.

5. Acknowledgements

This research was supported by the JPSP 21st Century COE program on informatics research for development of knowledge society infrastructure.

6. References

[1] Shinya Fujie, Yasushi Ejiri, Hideaki Kikuchi and Tetsunori Kobayashi, Dialogue Robot with an Ability to Understand Para-Linguistic Information, 2003-SLP-48, 13–20, 2003.

[2] David Sadek, Design Considerations on Dialogue Systems: From Theory to Technology -The Case of Artimis-, Proc. of ESCA IDS'99 Workshop, 173–187, 1999.

[3] Yukiko I. Nakano and Tsuneaki Kato, Utterance Content and Dialogue Strategies in Instruction Dialogue: Effects of the Discourse History and the Understanding Level of the Novice, SIG-SLUD-9502-4, 24–31, 1995.

[4] Senko K. Maynard, Japanese Communication - Language and Thought in Context, Univ. of Hawaii Press, Honolulu, 1997.

[5] Akira Ichikawa and Shinji Sato, Roles of Prosodies in Dialogue, 94-SLP-2, 51–58, 1994.

[6] Yohei Okato, Keiji Kato, Mikio Yamamoto and Shuichi Itahashi, Prosodic pattern recognition for insertion of interjectory responses and its evaluation, 96-SLP-10, 33–38, 1996.

[7] Hanae Koiso, Yasuo Horiuchi, Syun Tutiya and Akira Ichikawa, The acoustic properties of "subutterance units" and their relevance to the corresponding follow-up interjections in Japanese, SIG-J-9501-2, 9–14, 1995.