Extracting Interval Distribution of Human Interactions

Ryohei Kimura[1], Noriko Takemura[1], Yoshio Iwai[2], and Kosuke Sato[1]

[1] Graduate School of Engineering Science, Osaka University
{kimura,takemura}@sens.sys.es.osaka-u.ac.jp
Sato@sys.es.osaka-u.ac.jp
[2] Graduate School of Engineering, Tottori University
Iwai@ike.tottori-u.ac.jp

Abstract. Recently, activity support systems that enable dialogue with humans have been intensively studied owing to the development of various sensors and recognition technologies. In order to enable a smooth dialogue between a system and a human user, we need to clarify the rules of dialogue, including how utterances and motions are interpreted among human users. In conventional study on dialogue analysis, duration between the time when someone finishes an utterance and the time when another human starts the next utterance were analyzed. In a real dialogue between humans, however, there are sufficient intervals between an utterance and a visually observable motion such as bowing and establishing eye-contact; the facilitation of communication and cooperation seem to depend on these intervals. In our study, we analyze interactions that involve utterances and motions at a reception scenario by resolving motions into motion primitives (a basic unit of motion). We also analyze the timing of utterances and motions in order to structure dialogue behaviors. Our result suggest that a structural representation of interaction can be useful for improving the ability of activity support systems to interact and support human dialogue.

Keywords: Structural Representation of Interaction, Action Primitives, Interval Analysis.

1 Introduction

Recently, the development of various sensors and image and speech recognition technologies have fostered more intensive studies of systems that support dialogue with human [1,2,3]. Although interaction among humans or between a human and a system are diversifying, as in the systems mentioned above, it is difficult for humans to communicate with these systems as smoothly as they would in a daily face-to-face human conversation. For example, when humans interpret the thoughts of a dialogue partner, they holistically use not only aural information such as the context of utterance but also visual information such as body language including a nod, a gesture, or lip movements. In addition, during human conversation, it is important for one partner to start talking before or

Y.-S. Ho (Ed.): PSIVT 2011, Part II, LNCS 7088, pp. 262–273, 2011.

soon after the other partner finishes his or her utterance in order to communicate that the intent of utterance was understood. In conventional systems, the information for predetermined demands from users are communicated to other users and they cannot adjust them [4,5]. Redundant intervals between utterances occur because of delays in communication and low computational speed, and it becomes difficult to communicate the intent of the utterance correctly. In such situations, incorrect recognition of the intent of the utterance or a delay in the response from a dialogue partner can cause anxiety, because the partners cannot understand the intention and the utterances from each partner sometimes collide owing to redundant utterances. Therefore, in conventional study on dialogue analysis, various human interactions were analyzed to clarify the relationship between a speaker's intent and the structure of the interaction. However, tagging of speech intentions or evaluating temporal structures is done by one of the interacting participants or a third party and depends heavily on their subjective judgments. Therefore, in our study, we extract structural representations automatically from speech and visual information such as body movements during human interaction.

The rest of the paper is organized as follows. Section 2 presents a comparative analysis between our study and related studies, and section 3 describes the recognition model of conversation behavior. Section 4 describes the database of interaction behavior complied in a register that is used at the reception in the experiment, and section 5 describes our experimental results showing extracted response time distribution. Section 6 summarizes the paper and discusses future works.

2 Related Works

Kawashima et al. analyzed the temporal structure of head movements and utterances in *Rakugo* and extracted the timing when the performers switch roles by using visible gestures. Although they analyzed the temporal structure of multi modal interactions, such transitions are not represented structurally [9].

Sumi et al. proposed a structural representation of human interactions by using multi modal data obtained by using various sensors; this approach is similar to the used in the present study [10]. In that study, a three-party conversation in a poster presentation was represented by three annotations: utterance information, eye directions, and pointing directions. The study assumed that the transitions among speakers in the three-party conversation were affected by only the previous state and could be represented by a tree structure using an n-gram model. However, a semantic-level representation such as an annotation depends on the interpretation of individuals and requires manual extraction from sensor data. It is, therefore, difficult to extract a semantic-level representation automatically. Moreover, an n-gram model cannot express the duration of the dialogue, which plays an important role during an interaction, because the symbols used in an n-gram model express only the state of the dialogue.

In this paper, we model dialogue behavior by using action primitives, a minimal unit of action, which can be automatically extracted at the signal level from

sensor data, such as three-dimensional (3-D) joint positions or voice signals, and do not require manual annotations. The temporal relationship between dialogue partners is represented by the transition probability of dialogues learned from the intervals of the action primitives extracted from sensor data that were used as training data. Since a dialogue is modeled at a signal level and not at a semantic level, we can make a representation of the implicit information of dialogue partners, which cannot be achieved at a semantic-level.

3 Dialogue Recognition Model

In this section, the recognition model of dialogue behavior between humans are described.

3.1 Structuring the Dialogue Behavior

Fig. 1 shows how the model of dialogue behavior was structured for using in the present study. In our study, a dialogue behavior is represented by a motion primitive, m, which is a minimum unit of motion extracted automatically from multi modal data. The detailed method used to extract motion primitives is described in Sec.4. Extracted motion primitives are denoted by $m_1, m_2, ..., m_N$ in the order of observation, and the start time and finish time of m_i are denoted by t_s^i and t_e^i respectively. We define a basic dialogue behavior I_S as an interactional subsequence, i.e., a prior motion primitive of an asker $m_a \rightarrow$ recognizing m_a by a responder $\rightarrow$ a posterior motion primitive of the responder $m_r \rightarrow$ recognizing m_r by the asker. Sequential dialogue behavior is expressed as the transition of the basic dialogue behavior. The posterior motion primitive m_r^i in the i-th basic dialogue behavior I_S^i is equivalent to the prior motion primitive m_a^{i+1} in the $i+1$-th basic dialogue behavior I_S^{i+1}. Human dialogue behaviors are defined as a chain model of following three probabilities: p_0, p_1, and p_2. For illustrative purposes, it is assumed that human A is an asker and human B is a responder in the basic dialogue behavior I_S^i.

- p_0: the probability of an initial motion primitive
  p_0 is the probability of observing the motion primitive m_a^1 by human A when no motion primitives were observed right before that time. p_0 is defined as follows:
  $$p_0^A(m_a^1) \tag{1}$$
- p_1: the probability of transiting motion primitives ($m_a \rightarrow m_r$)
  p_1 is the probability of starting the motion primitive m_r^i by human B at time t_s^{i+1} when the motion primitive m_a^i finish at time t_e^i in the basic dialogue behavior I_S^i. p_1 is defined as follows:
  $$p_i^B(m_r^i, t_s^{i+1} | m_a^i, t_e^i) \tag{2}$$

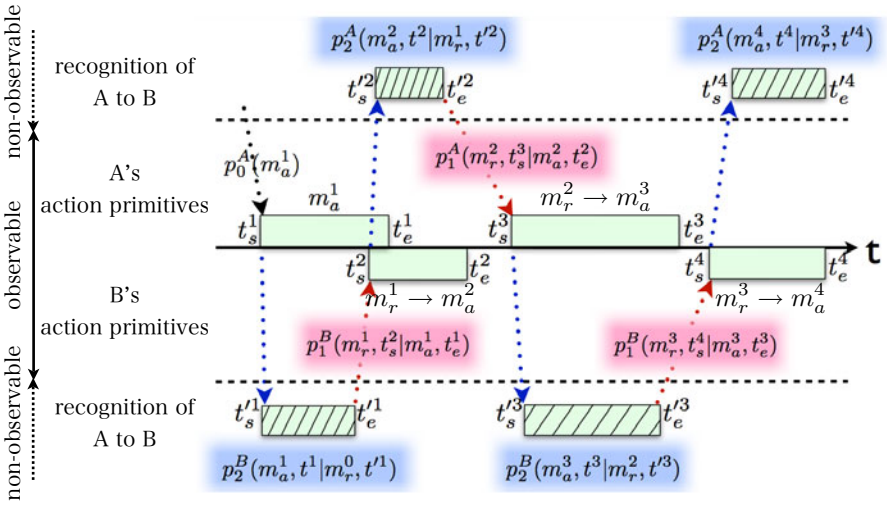

Fig. 1. Structure of the dialogue behavior model

– p_2: the probability of transiting basic dialogue behaviors ($I_S^i \rightarrow I_S^{i+1}$)
When the basic dialogue behavior I_S^i transits to I_s^{i+1}, the posterior motion primitive m_r^i by human B in I_S^i becomes the prior motion primitive m_r^{i+1} in the next basic behavior I_S^{i+1}. However, it does not always occur. This transition occurs only when human A needs to respond to a motion by human B. The probability p_2 of regarding the response m_r^i at the current step as the prior motion m_a^{i+1} at the next step is calculated as:

$$p_2^A(m_a^{i+1}, t^{i+1} | m_r^i, t'^{i+1}) \tag{3}$$

where t'^i denotes the recognition time against m_a^{i+1}. If the difference between the motion primitive by the asker and its recognition by the responder is large, the dialogue sometimes breaks. The probability seems to depend on invisible factors, such as the mental states, intentions, social status of each human, and the environment in which the dialogue occurs.

By using these probabilities, the observed dialogue behavior can be described by following a probability chain model:

$$p^A(m_a^1) p_2^B(m_a^1, t'^1 | m_r^0, t^1) p_1^B(m_r^1, t_s^2 | m_a^1, t_e^1) p_2^A(m_a^2, t'^2 | m_r^1, t^2) \cdots p_2^B(m_r^N, t'^N | m_a^{N-1}, t^N) \tag{4}$$

The purpose of our study is to extract probability p_1 in the proposed model. By using the dialogue data while performing the same task as the one performed in the experiment for estimating p_2, p_2 can be approximated to 1 except for the end time of motion primitive.

4 Extraction of Response Time Distribution

In this section, the method used to learn the transition probability of dialogue behaviors is described.

4.1 Extraction of Motion Primitive

In our study, body motions in dialogues are represented by motion primitives, which are the smallest units of motion. A motion primitive is similar to a phoneme used for phonetic recognition. Motion is described as a transition of motion primitives, i.e., typical postures. The advantage of our method is that it is less subject to individual differences in terms of physical features, because only the information obtained from postures is used to define motions. Another advantage is that the method does not require manual procedures, such as annotating postures. The method to extract motion primitives is described as follows.

Preprocessing
First, 3-D position data $\mathbf{x} = (x, y, z)^t$ of specific sites on the body of a speaker is measured at 30 fps by Microsoft's Kinect system using a range sensor and a camera. Eight positions on the body are measured: the head, the chest, both shoulders, both elbows, and both hands, as shown in Fig.2.

When the measurement of a 3-D position fails, the position is estimated by a linear interpolation between the previous point and the next point at the failure point. We used linear interpolation because the duration of failure were relatively short in the experiment and the movements of body site positions in the short term could be estimated by a uniform linear motion model.

The 3-D positions of the body sites in an absolute coordinate system were obtained using the Kinect system: however, it is necessary to convert coordinates in order to recognize that the postures are equivalent to the ones that have the same features but face in another direction or position, i.e., it is necessary to convert coordinates for posture recognition without depending on the direction and position of the posture. To overcome this problem, the coordinate origin is shifted to the center of the torso region and the body direction is normalized. This makes it possible to observe the relative motion as seen from the center of the torso region and to extract the motion primitive without depending on the human's position and direction.

Classification of Postures
The motion primitives are extracted from the obtained 3-D position data. First, joint angles of the upper body $\mathbf{\Theta} = \{\theta_i | i = 1, 2, ..., 8\}$ are obtained as a feature of the posture from the 3-D position data. The advantages of using joint angle information are as follows:

- The number of feature points is low.
- There is a poor correlation between each feature point.
- The posture recognition is less subject to individual differences.

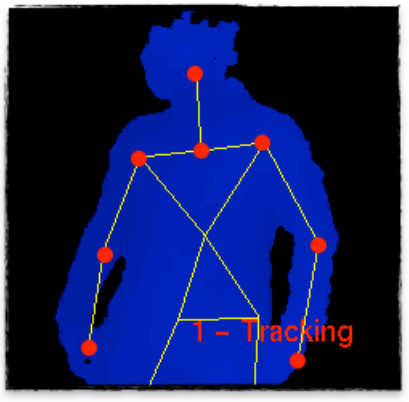

Fig. 2. The 3-D points of joints (red circles) measured by the Kinect system

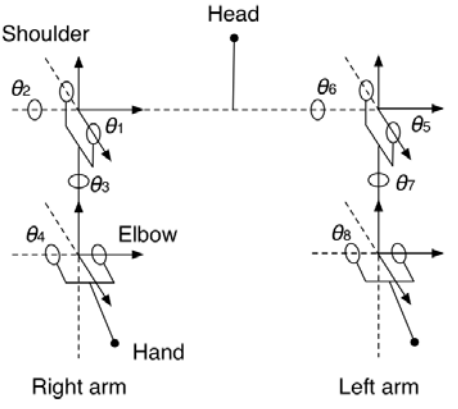

Fig. 3. Posture parameters

Each cluster of posture features generated by a clustering method based on a k-means algorithm is regarded as a motion primitive. The number of clusters is equivalent to the number of motion primitives and the postures in the same cluster are classified as the same motion primitive. Optimal k is estimated through experiments.

By this procedure, the transition of postures with joint angles are obtained as time-series data from each set of learning data.

4.2 Interval Detection of Utterances

Intervals of utterances are detected from speech waves obtained by a pin microphone attached to the speaker's clothes. In the detection method, the power of a speech wave is calculated by Fast Fourier Transform (FFT) with a divided speech wave and converted to a binary form using an adequate threshold. We interpolated using an adequate time threshold between two nearby intervals of utterances in order to deal with the case when the detection of an utterance interval fails owing to uneven voice volume.

4.3 Detection of Response Time Distribution

Based on the obtained information about motion primitive transitions, values for calculating the probabilities (p_1) of prior motion, posterior motion, and response time between prior motion and posterior motion were obtained. Response time is the time between the end of one (prior) motion and the start of next (posterior) motion. We extracted response times from a large amount of learning data and plotted histograms in order to estimate the shape of its probabilistic distribution. The accuracy of parametric estimation can be improved by iterative learning with estimated probabilistic distribution.

5 Experimental Results

5.1 Experimental Conditions

In this study, we conducted experiments to extract the response time distribution of interaction behaviors. We assumed the situation a wedding reception scenario and extracted the response time of dialogues and behaviors between a guest and a receptionist. In this situation, the start and end times of behavioral actions are clear, and many interactions such as offering greetings, hand-delivering wedding presents, and signing the guest book are performed at the reception. The behavioral data were collected by using the method described in the previous section and the layout of sensors is shown in Fig. 4. The red dashed line area in Fig. 4 shows the area where motion was captured by Kinect sensors. A guest book and a pen were placed on the table. The relative 3-D positions of joints and the time of utterance were stored as exemplars in a behavior database. The duration of an interaction at the reception area is about 1 min, and no restrictions are given to subjects except the scenario shown in Figs. 5, 6, 7 and 8.

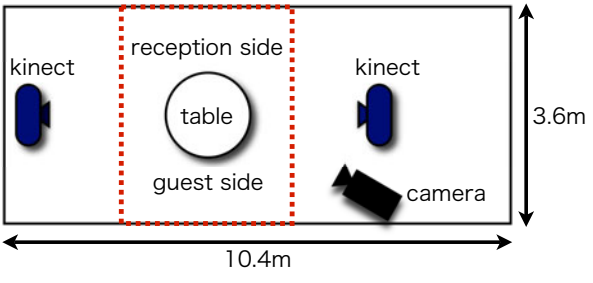

Fig. 4. Arrangement of the experimental environment

5.2 Extraction of Response Time Distribution

From the behavior database, we extracted motion primitives of behavior and the distribution of the response times between motion primitives. It is difficult to determine the number of motion primitives because if the number is too small we cannot capture the structure of the interaction, and if the number is too large we cannot extract meaningful distributions. In this study, we determined the number of action primitives to be six through the preliminary experiment. The extracted response time distribution is shown in Fig. 10 when the number of action primitives is six. Fig. 10 shows the histogram of each pair of motion primitives. As shown in Fig. 9, the horizontal axis shows the interval time for each 200 ms when pre-action primitive m^3 is completed and the vertical axis shows the frequency of the post-action primitive m^5. A negative interval time means an overlap of action primitives.

When extracting the response time, we ignored the action primitives when the response time was smaller than -1200 ms or greater than 1400 ms, because the relationship between such action primitives is weak. The distribution shapes of the response times are divided into four types:

Fig. 5. Arrival of the guest

Fig. 6. Hand-delivery of presents

Fig. 7. Signing the guest book

Fig. 8. Guest leaves

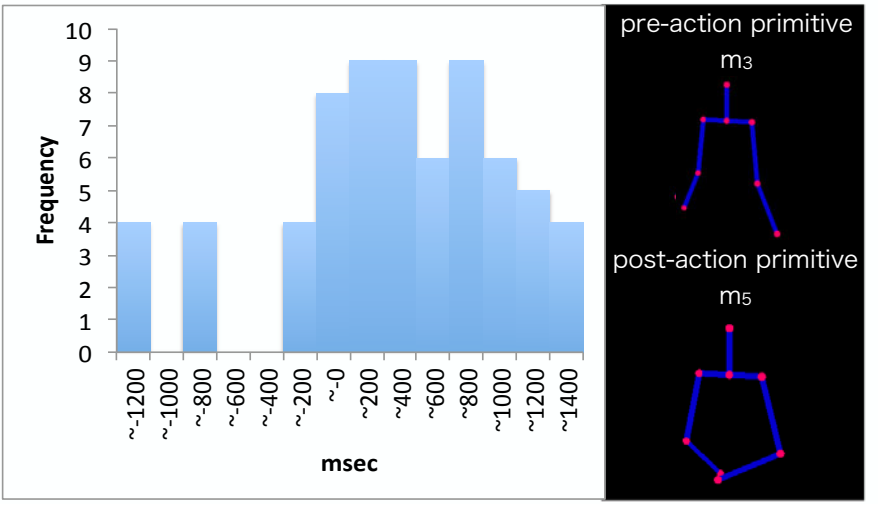

Fig. 9. The histogram of response time (pre-action primitive m^3 and post action primitive m^5)

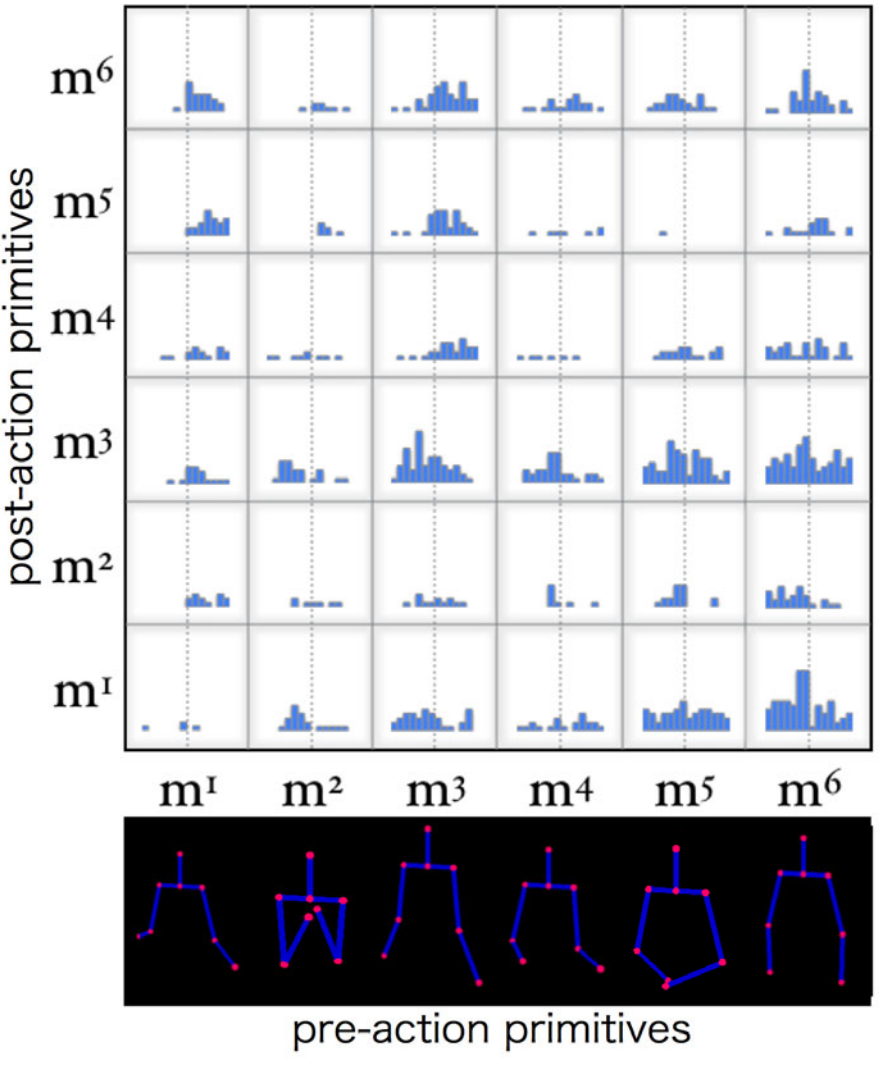

Fig. 10. Extraction of the response times when the number of action primitives is six

(i) The distribution has a peak between the termination time of the preceding action and 1000 ms after the preceding action. (e.g., a pair of (m^3, m^6)),
(ii) The distribution has a peak between -600 ms and the termination time of the preceding action (e.g., pairs of (m^5, m^3) and $(m^6 and m^1)$),
(iii) The distribution has no explicit peak (e.g., a pair of (m^5, m^1)),
(iv) There is no correlation between action primitives (e.g., pairs of (m^2, m^4) and (m^4, m^5)).

In types (i) and (ii), the frequency of the following action primitive increases just after the preceding action primitives or when they overlap. There might be

a distribution of the response time with a peak same as that for an utterance. For type (iii), the results cannot be treated as a distribution with an explicit peak. The reasons of this phenomenon are lack of training data, in accuracy of action primitive extraction, or no correlation between action primitives in real time. In type (iv), the correlation between action primitives is very low, so the transition probability of such an interaction is very low.

5.3 Distribution Extraction under Utterance Effect

Next, we conducted an experiment to investigate the distribution changes affected by utterances. In this experiment, the post-action primitives are extracted only when an utterance is detected, and then compared with the results in Fig. 10. Figure 11 shows the distribution of the response time of the post-action primitives extracted only when an utterance is detected. Comparing Fig. 11 with Fig. 9, we observe that the peak of the distribution of the post action primitives is shifted forward. The reason for this shift is that the post-action primitives are suspended until the guest or receptionist finishes his or her utterance. From these results, it can be concluded that there exists a valid response time between the preceding action primitives and post action primitives, but the distribution of the response time changes when the preceding action primitives are performed with an utterance. Therefore, when activity support systems return a response, they are effective for creating visual and acoustic reply patterns.

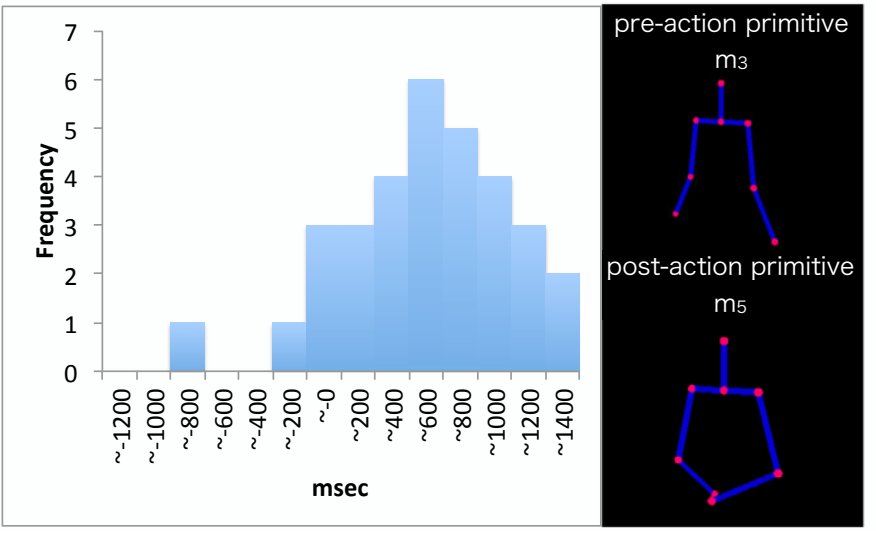

Fig. 11. The histogram of the response time when an utterance is detected (pre-action primitive m^3 and post-action primitive m^5)

6 Conclusion and Future work

In this study, we extracted the distribution of the response time of utterances and action primitives by detecting 3-D positions of joints and utterances. We also proposed a method to structurally model dialogues. In order to show the effectiveness of the response time between action primitives in s dialogue, we collected dialogue data from a wedding reception scenario and analyzed the distribution of the response times. From this analysis, we found that an appropriate interval existed for a response time of action primitives in the dialogue, and that the interval is affected and differs by a pair of action primitives and the presence of an utterance.

In future work, we will investigate a method for recognizing multi modal dialogue by using structural representation of human interactions while maintaining an appropriate interval between action primitives. This will lead to the realization of a system that can enable dialogue between users, which appear more natural, by using the structural representation of action primitives for generating responses from the system.

Acknowledgement. This work is partially supported by cooperative research with Daiwa House Industry Co., Ltd. and by Grant-in-Aid for Scientific Research on Innovative Areas (No. 22118506).

References

1. Takizawa, M., Makihara, Y., Shimada, N., Miura, J., Shirai, Y.: A Service Robot with Interactive Vision-Object Recognition Using Dialog with User. In: Proc. of the 1st Int. Workshop on Language Understanding and Agents for Real World Interaction (Academic Journal), pp. 16–23 (July)
2. Kuriyama, H., Murata, Y., Shibata, N., Yasumoto, K., Ito, M.: Congestion Alleviation Scheduling Technique for Car Drivers Based on Prediction of Future Congestion on Roads and Spots. In: Proc. of 10th IEEE Int'l. Conf. on Intelligent Transportation Systems (ITSC 2007), pp. 910–915 (September 2007)
3. Fukaya, K., Watanabe, A.: Intuitive Manipulation to Mobile Robot by Hand Gesture. In: 24th ISPE International Conference on CAD/CAM, Robotics and Factories of the Future (July 2008)
4. Shirberg, E.: Spontaneous Speech: How People Really Talk and Why Engineers Should Care. In: Proc. EUROSPEECH (2005)
5. Fujie, S., Fukushima, K., Kobayashi, T.: Back-channel Feedback Generation Using Linguistic and Nonlinguistic Information and Its Application to Spoken Dialogue System. In: Proc. EUROSPEECH, pp. 889–892 (2005)
6. Hirose, K., Sato, K., Minematsu, N.: Emotional speech synthesis with corpus-based generation of F0 contours using generation process model. In: Proceedings of International Conference on Speech Prosody, Nara, pp. 417–420 (March 2004)
7. Fujiwara, N., Itoh, T., Araki, K.: Analysis of Changes in Dialogue Rhythm Due to Dialogue Acts in Task-Oriented Dialogues. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 564–573. Springer, Heidelberg (2007)

8. Nishimura, R., Kitaoka, N., Nakagawa, S.: Analysis of relationship between impression of human to human conversations and prosodic change and its modeling. In: Proceeding of the Interspeech, pp. 534–537 (2008)
9. Kawashima, H., Nishikawa, T., Matsuyama, R.: Analysis of Visual Timing Structure in *Rakugo* Turn-taking (written in Japanese). IPSJ Journal 48(12), 3715–3728 (2007)
10. Sumi, Y., Yano, M., Nishida, T.: Analysis environment of conversational structure with nonverbal multimodal data. In: 12th International Conference on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010), Beijing, China (November 2010)