Modeling User's State During Dialog Turn Using HMM For Multi-modal Spoken Dialog System

Yuya Chiba, Akinori Ito
Graduate School of Engineering,
Tohoku University
Email: {yuya, aito}@spcom.ecei.tohoku.ac.jp

Masashi Ito
Department of Electronics and Intelligent System,
Tohoku Institute of Technology
Email: itojin@totec.ac.jp

Abstract—**Conventional spoken dialog systems cannot estimate the user's state while waiting for an input from the user because the estimation process is triggered by observing the user's utterance. This is a problem when, for some reason, the user cannot make an input utterance in response to the system's prompt. To help these users before they give up, the system should handle the requests expressed by them unconsciously. Based on this assumption, we have examined a method to estimate the state of a user before making an utterance by using the non-verbal behavior of the user. The present paper proposes an automatic discrimination method by using time sequential non-verbal information of the user. In this method, the user's internal state is estimated using multi-modal information such as speech, facial expression and gaze, modeled using a Hidden Markov Model (HMM).**

Keywords- multi-modal information processing; user's state; spoken dialog system

I. Introduction

Most spoken dialog systems estimate the user's internal state to generate an appropriate response to the user. Many researches on user modeling have been conducted such as emotion [1], [2], preference [3] and familiarity with the system [4]. These methods implicitly assume that the user always gives some responses to the system's prompt. However, not all users can use the system proficiently. For instance, a user may abandon a session without uttering a word if he or she cannot understand the meaning of the system's prompt, or could take a long time to consider how to answer the prompt. The user model which does not depend on linguistic information is needed to help these users appropriately. To tackle this problem, we have assumed two basic internal states of a user who cannot make an utterance. The first one is the state where the user does not know what to input, and the second one is when the user is considering how to answer the system's prompt. We call these states "state A" and "state B", respectively. Although our definitive goal is building a spoken dialog system that can help the user in an optimum manner, we focus on the estimation method of user's state in this report. Since discrimination must be processed after the user gets the floor and before he/she makes the input utterance to the system, we denote them as the user's internal state "during a dialog turn." Figure 1 shows typical examples of these user's states.

In a human-human dialog, interlocutors converse while more or less estimating the internal state of the dialog partner based on the feeling that the other person knows the answer to the question (in other words, whether other interlocutors could respond to his/her utterance or not). This ability is called "Feeling of Another's Knowing" (FOAK) [5][6]. It is thought that introducing such manner of human dialog into a dialog system would improve its performance.

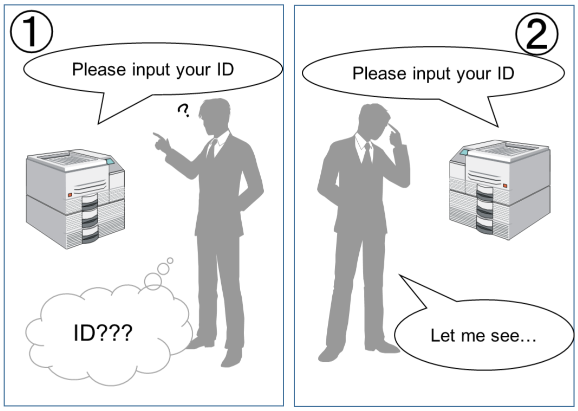

Figure 1. Target user's state

To estimate the user's state during a dialog turn, we used visual and acoustic features instead of linguistic contents of the utterance. Bi-modal feature fusion has been examined in the field of emotion recognition [7][8]. In the present research, almost the same features as these studies were employed because the user's state during a dialog turn has similar aspects to emotion. Besides, successive estimation is required to achieve the goal of our work because there is no explicit trigger for processing the user's utterance.

In the previous report [9], we used a multi-stage neural network to integrate multi-modal features and estimate the user's state frame by frame. However, this method has the problem that it cannot capture the temporal variation of the features. Based on this result, the present study examines an estimation method which uses multi-stream Hidden Markov Model (HMM) to model the local temporal variation of the audio-visual feature sequence. In this method, the likelihoods of the user's states are obtained continuously. The performance of the proposed method is evaluated by discrimination examination.

This paper is organized as follows. Collection of the experimental data is described in Section 2. Then estimation method

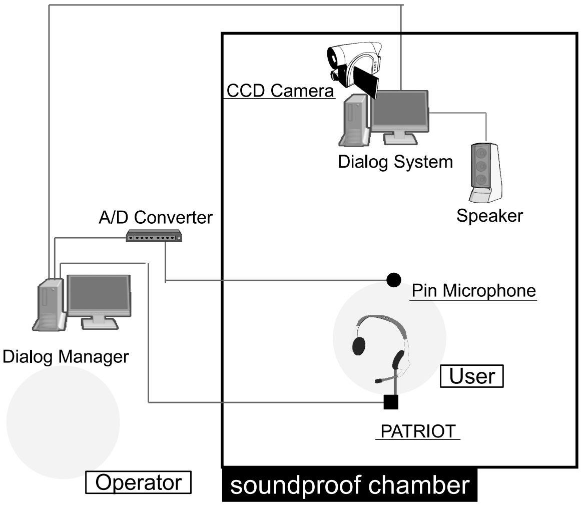

Figure 2. Experimental circumstance

TABLE I. EVALUATION RESULTS

State A	State B	Neutral	Total
59	195	538	792

using multi-stream HMM is introduced in Section 3. The multi-modal features employed in this paper is described in Section 4. Finally, the results of the experiments are presented in Section 5.

II. EXPERIMENTAL DATA

Experimental data were collected on the Wizard of Oz basis. The dialog experiments were conducted in a soundproof chamber. We implemented a question-and-answer task in which the system posed questions and the subjects answered them. The task was designed to make the user as embarrassed as possible. The questions were about common knowledge or a number memorized in advance, such as "Please input current date." and "Please input your ID number." Additionally, an agent with a simple cartoon-like face was projected on the monitor to keep the subjects' attention. Figure 2 shows an experimental circumstance. We employed 16 subjects (14 males and 2 females). The subjects wore a lapel microphone. To record an image of the subjects' frontal face, a CCD camera was installed above the monitor in front of the subjects. The operator remained outside of the chamber and controlled the agent remotely. The audio signal was recorded in PCM format at 16 kHz sampling, 16-bit quantization. The recorded video clips were stored as AVI files with 24-bit color depth, 30 frame/s. After the experiment, we separated the dialog into sessions; one session included one interchange of the system's prompt and the user's response. Here, we defined the length of the segment between the end of the system's prompt and the beginning of the user's input utterance as "latency". Sessions with more than 5.0 s latency were labeled by five evaluators.

Table I shows the results of the evaluation. The label of each session was chosen by majority vote of evaluators (Fleiss' $\kappa = 0.22$). One session was excluded because the acoustic feature could not be extracted due to overlapping utterances.

III. DISCRIMINATION METHOD USING MULTI-STREAM HMM

The spoken dialog system has to detect whether the user needs help or not as soon as possible, because the ultimate purpose of our work is to build a system that responds to a user who has difficulty in answering a question. Therefore, we need incremental evaluation of the user's internal state, and the system should help the user just after detecting the stagnation of the dialog. We therefore observed the time sequence of the features of the user, and fed the features to the classifier frame by frame. We assume the user's non-verbal behavior was recorded continuously during the dialog with the microphone and the CCD camera. In the previous paper, we have examined the method to estimate the user's state by using a single audio-visual feature frame as a feature vector of the classification. However, capturing the temporal variation of the feature sequence is considered to be essential for the estimation. For instance, the user thinking of the response to the system (i.e., state B) tends to emit long fillers, and the user cannot understand the meaning of the prompt (i.e., state A) might move his/her eyes frequently. Therefore, we proposed the method that extracts the segment of the feature frames as a vector and feeds the feature vector to the HMM in order to represent temporal characteristics of the feature sequence. Additionally, we used a multi-stream HMM as the classifier. The multi-stream HMM can fit the distribution of the multi-modal features efficiently by dealing with the feature sequence belonging to the different modality as the different stream. The score of the user's state is emitted frame by frame, and they are integrated at the final stage to decide the discrimination result. The topology of the HMM was ergodic; therefore, the HMM has transitions between all states.

The output probability of the state j, $b_j(\boldsymbol{o}_t)$ is denoted as follows:

$$b_j(\boldsymbol{o}_t) = \prod_{s=1}^{S} \left[\sum_{m}^{M_s} c_{jsm} \mathcal{N}(\boldsymbol{o}_{st}; \boldsymbol{\mu}_{jsm}, \boldsymbol{\Sigma}_{jsm}) \right]^{\lambda_s} \tag{1}$$

where, $\boldsymbol{o}_{st}$ is the feature sequence belonging to stream s at time t, c_{jsm}, μ_{jsm}, Σ_{jsm} is the parameter of the output probability density function of the state j and represents weights of mixture, mean vector, and covariance matrix, respectively. Each stream corresponds to each modality (indicated in the next section). $\mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate Gaussian function, that is;

$$\mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{o}-\boldsymbol{\mu})^t \Sigma^{-1}(\boldsymbol{o}-\boldsymbol{\mu})\} \tag{2}$$

These equations show that the multi-stream HMM emits the output log likelihood as the weighted sum of the output log likelihoods of each stream.

In addition, we show the overview of the construction of the segmental feature in Figure 3. Segmental feature f_t is constructed to have the past n frames from time t and obtained by shifting one frame at a time. Therefore, the segmental feature enables both the investigation of temporal characteristic of the feature sequence of the short segment and frame by frame estimation of the user's state. As shown in Figure 3 the number of frame n was set to 100, which is equivalent to one second in real time.

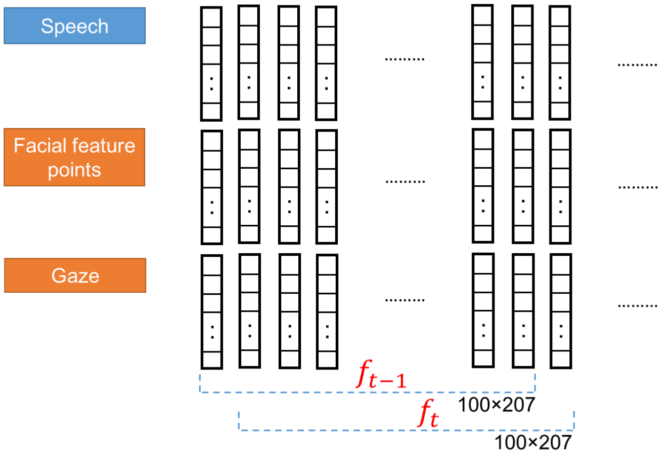

Figure 3. Segmental feature construction

TABLE II. CONDITIONS OF ACOUSTIC FEATURE EXTRACTION

	MFCC	ΔPitch	Zero cross ratio
Frame width	25.0 ms	7.5 ms	10.0 ms
Frame shift	10.0 ms	10.0 ms	10.0 ms

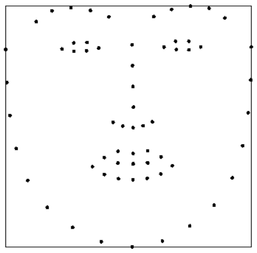

Figure 4. Model of facial feature points

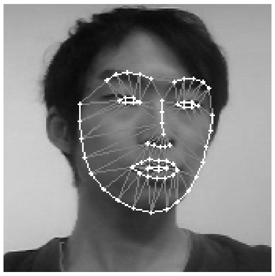

Figure 5. Result of feature extraction

IV. MULTI-MODAL SEQUENTIAL FEATURES

The target user's states are assumed to have similar aspects to emotion, and therefore we employed almost the same features used in the area of emotion recognition, such as the spectral features of the speech, intonation, zero cross ratio, facial feature points and gaze direction. In particular, it is suggested that emotion has a multi-modality nature [10], and most researches reported that recognition accuracy is improved by combining multi-modal information [7], [8].

A. *Acoustic feature*

To represent spectral characteristics, Mel-Frequency Cepstrum Coefficients (MFCC) is employed as a low-level acoustic feature. In our method, velocity and acceleration coefficients of MFCC including the log power were used. The total number of dimensions of MFCC is 39 and the frame length of calculating the time difference components is 5. Intonation of speech is represented by an $F0$ contour. The $F0$ was extracted by the normalized cross correlation, then converted to the log-scale. Since the $F0$ has large variation from speaker to speaker, a differential coefficient is used as the acoustic feature. In addition to the features mentioned above, the zero cross ratio is used to distinguish voiced and unvoiced segments. The basic conditions for extracting each acoustic feature are shown in Table II.

B. *Facial feature points*

Facial activity of the user is the most important feature among the visual information. To represent the facial activity, feature points of the face were extracted by Constraint Local Model (CLM) [11]. In this method, a model of the feature points is fitted after detecting the facial region from the whole image in the frame. Figure 4 shows a model of feature points and Figure 5 is an example of the result of fitting. The fitting error is mainly caused by misdetection of the facial region and occlusion. Although misdetection of the facial region was corrected by hand for the examination, the error caused by occlusion was not considered in the feature extraction; 5% of all sessions contained severe fitting errors. We used the relative coordinates of the feature points as the visual features. The number of feature points was 66 and the number of dimensions of the features was 132. The locations of feature points were normalized by the size of the facial region.

C. *Gaze feature*

Previous analysis showed the user's gaze action affects the evaluation of the user's state. Therefore, we used brightness feature of the eye region to represent the broad location of the user's iris indirectly. We employed Haar-like feature which has the fast calculation algorithm using the integral image. The Haar-like feature is extracted by applying filters depicted as Figure 6 to the image and originally used for object detection. We calculated the Haar-like feature from both eyes region obtained by CLM. As the Haar-like feature vector has high dimensionality, the principal component analysis (PCA) is applied for reducing the dimensionality of the feature vectors. After reducing the dimensions, the gaze feature has 34 dimension and cumulative contribution rate was about 95 %.

D. *Feature synchronization*

Finally, these features were synchronized because audio and visual information were extracted by different sampling rate. We synchronized the features by copying the previous visual feature values for each 10 ms. Therefore, the number of dimensions of a combined feature frame was always 207.

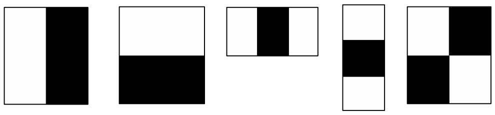

Figure 6. Haar-like filter

TABLE III. EXPERIMENTAL CONDITION

Experimental data	State A(59), State B(195)
Number of states	3, 4, 5
Number of mixture components of each stream	4

V. DISCRIMINATION EXAMINATION

A. Experimental condition

To evaluate the proposed method, we conducted a discrimination examination. Table. III shows the experimental condition. The number in parentheses indicates the number of sessions.

The previous work [9] showed that discrimination between the neutral state and the other states can be done by using the latency. Therefore, we focus on the discrimination between the state A and state B in the following experiment. Although the weights of streams should be decided to optimize the discrimination results, we fixed all stream weights to 1.0 to verify the effectiveness of estimation using HMM. The optimization method of the stream weights is a future work. In this paper, we changed the number of states of the HMM and evaluated the discrimination accuracy.

The definitive discrimination result was decided by comparing the average scores of each states. That is:

$$\hat{c} = \arg\max_{c} \frac{1}{T} \sum_{t=1}^{T} p_{tc} \tag{3}$$

where T is the length of the segment for which the state is estimated. In this experiment, we used the duration of each session as T. Here, the total accuracy tends to increase as the determined class leans toward state B because the amount of data is not uniformly distributed (see Table. I and Table. III); therefore, the harmonic mean (denoted as $Harm.$) was employed for measuring the performance. This is calculated by

$$Harm. = \frac{2 \cdot Corr_A \cdot Corr_B}{Corr_A + Corr_B} \times 100.0 \quad (\%) \tag{4}$$

where $Corr_A$ and $Corr_B$ represent the discrimination accuracy of state A and state B, respectively. The experiments were conducted by 5-fold cross validation.

B. Experimental result

Figure 7 shows the experimental results. We showed the result of the previous experiment [9] (Baseline in Figure 7) for comparison with the proposed method. The best performance was obtained when the number of states of HMM was 4 ($Harm. = 64.4\%$). This result was about 2 points higher than the baseline method, which used a neural network and a single feature frame. However, the improvement of the performance was not as large as expected, considering the additional computation cost. One of the reasons is the output distributions of each stream are not learned enough due to the shortage of the training data. The multi-stream HMM has an advantage in that it can control the importance of the output probability of each stream by changing the stream weights. Therefore, we will examine the method to optimize the stream weights to improve the performance in a future work.

VI. CONCLUSION

We examined the user modeling of dialog turn using a multi-stream HMM and segmental feature to represent temporal variation of an audio-visual feature sequence. The proposed method obtained the best result when the number of state was 4, and the result surpassed our previous work. On the other hand, we also observed the limitation of the discrimination performance of multi-stream HMM without optimizing the stream weights. In a future work, we will examine the method to decide the stream weights to improve the discrimination result.

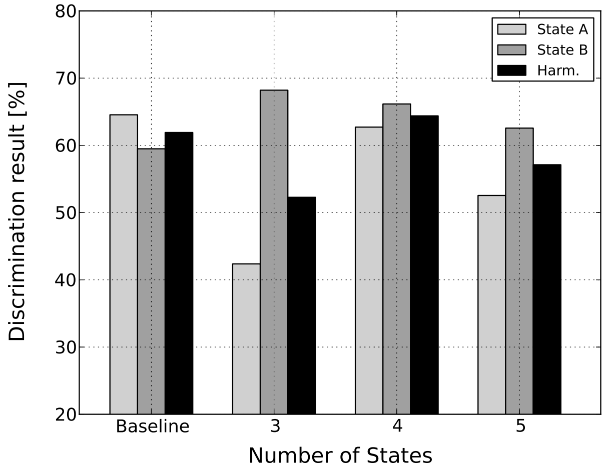

Figure 7. Discrimination results

REFERENCES

[1] K. Forbes-Riley and D. Litman, “Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor,” Speech Communication, vol. 53, 2011, pp. 1115–1136.

[2] A. Metallinou *et al.* “Context-sensitive learning for enhanced audiovisual emotion classification,” IEEE Trans. Affective Computing, vol. 3, no. 2, 2012, pp. 184–198.

[3] A. N. Pargellis, H. K. J. Kuo, and C. H. Lee, “An automatic dialogue generation platform for personalized dialogue applications,” Speech Communication, vol. 42, 2004, pp. 329–351.

[4] K. Jokinen and K. Kanto, “User expertise modelling and adaptivity in a speech-based e-mail system,” in Proc. COLING, 2004, pp. 87–94.

[5] S. E. Brennan and M. Williams, “The feeling of another’s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers,” J. Memory and Language, vol. 34, no. 3, 1995, pp. 383–398.

[6] M. Swerts and E. Krahmer, “Audiovisual prosody and feeling of knowing,” J. Memory and Language, vol. 53, no. 1, 2005, pp. 81–94.

[7] J. C. Lin, C. H. Wu, and W. L. Wei, “Error weighted semi-coupled hidden markov model for audio-visual emotion recognition,” IEEE Trans. Multimedia, vol. 14, no. 1, 2012, pp. 142–156.

[8] Y. Wang and A. N. Venetsanopoulos, “Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition,” IEEE Trans. Multimedia, vol. 14, no. 3, 2012, pp. 597–607.

[9] Y. Chiba, M. Ito, and A. Ito, “Estimation of userfs state during a dialog turn with sequential multi-modal features,” in HCI International 2013-Postersf Extended Abstracts, 2013, pp. pp. 572–576.

[10] O. Collignon *et al.* “Audio-visual integration of emotion expression,” Brain research, vol. 1242, 2008, pp. 126–135.

[11] J. M. Saragih, S. Lucey, and J. F. Cohn, “Deformable model fitting by regularized landmark mean-shift,” Int. J. Computer Vision, vol. 91, 2011, pp. 200–215.