Estimation of User's State during a Dialog Turn with Sequential Multi-modal Features

Yuya Chiba[1], Masashi Ito[2], and Akinori Ito[1]

[1] Tohoku University,
Aoba 6-6-5, Aramaki, Aoba-ku, Sendai, Miyagi
yuya@spcom.ecei.tohoku.ac.jp
[2] Tohoku Institute of Technology,
Kasumicho 35-1, Yagiyama, Taihaku-ku, Sendai, Miyagi

Abstract. Spoken dialog system (SDS) is a typical speech application and sometimes regarded as one of ideal interfaces. However, most of conventional SDSs cannot help their user while waiting for input utterance since they treat a user's utterance as a trigger of processing. This architecture is largely different from the manner of human-human interaction and factor that makes the user feel inconvenience when they cannot respond to the system's prompt appropriately. To solve this problem, the system should be able to estimate the internal state of the user before observing the user's input utterance. In present paper, we proposed two-step discrimination method using multi-modal information to estimate the user's state frame by frame.

Keywords: spoken dialog system, user modeling, multi-modal information.

1 Introduction

Spoken dialog systems need to estimate the user's internal state in order to generate an appropriate response to the user. Many researches have been conducted so far [1-4], but they implicitly assume that the user always makes some response to the system's prompts. 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. We therefore considered that two internal states of a user who cannot make an utterance should be taken into account. The first one is the state where the user does not know what to input, and the second one is that where the user is considering how to answer the system's prompt. We call them state A and state B, respectively. We also assume state C, where the user has no problem answering the system. The purpose of this study is to discriminate these three states. Since the discrimination must be processed before the user's input, we denote them as the user's internal state "during a dialog turn." In our previous study [5], we made an attempt to distinguish these three states using an SVM and the features extracted from the whole video sequences. It worked well, but we cannot use the whole sequence as an input for the

C. Stephanidis (Ed.): Posters, Part II, HCII 2013, CCIS 374, pp. 572–576, 2013.

purpose to help the user who has difficulty to make an utterance. The present study proposes an automatic estimation method: we employed a two-step neural network to model the audio and visual information, and obtained the results of the estimation frame by frame.

2 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 embarrassed as much as possible. The questions asked about common knowledge or a number memorized in advance. Additionally, an agent with a simple cartoon-like face was projected on the monitor to keep the subjects' attention. Figure 1 shows an experimental circumstance. We employed 16 subjects (14 males and 2 females). The subjects wore a lapel microphone. To record 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 1 shows the results of the evaluation. The label of each session was chosen by majority vote. The sessions shorter than 5.0 s were categorized as state C. One session was excluded because the acoustic feature could not be extracted due to overlapped utterances.

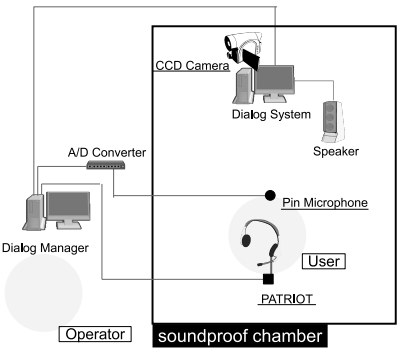

Fig. 1. Experimental circumstance

Table 1. Evaluation results

State A	State B	State C	Total
59	195	538	792

3 Discrimination Method

3.1 Hierarchical Discrimination

The system need incremental evaluation of the user's state to help just after detecting the user's embarrassment. Therefore, sequential features of the user are extracted and fed to the classifier frame by frame. The user's non-verbal behavior was recorded continuously during the dialog by the microphone and the CCD camera. Neural networks are used as the classifiers in the present paper. The front-end neural network outputs the scores of symbolic phenomena such as speech events or facial expressions. These outputs of the front-end network are used as the inputs of the back-end step. The back-end neural network outputs the definitive results, which are the scores of the states of the user.

3.2 Multi-modal Feature Selection

MFCC, fundamental frequency ($F0$) and zero cross ratio of the speech signal was employed as the low-level acoustic features. Here, MFCC contains their first and second derivatives and the total number of dimension was 39. $F0$ was calculated by cross-correlation method and first derivative was used for estimation. These audio features were extracted each 10.0 ms. The facial activity of the user is also important feature among the visual information. Therefore, feature points of the face were extracted by the method of Constraint Local Model (CLM) [6] and employed as the low-level visual feature. Figure 2 shows a model of feature points and Figure 3 is an example of the result of fitting. We used the relative coordinates of the feature points as the visual features. The number of feature points was 66 and the number of the dimensions of features was 132. The locations of feature points were normalized by the size of the facial region.

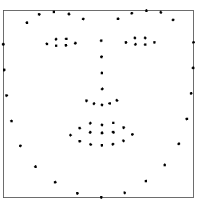

Fig. 2. Model of facial feature points

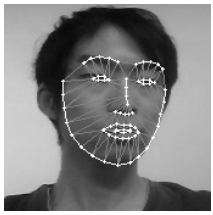

Fig. 3. Result of feature extraction

Table 2. Label of acoustic event

Acoustic events
System's prompt (AS)
Input utterance (AI)
Filler utterance (AF)
User's Aspiration or breath (AB)
Self speaking of the user (ASE)
Whisper of the user (AW)
Soundless segment (ASI)

Table 3. Label of visual events

Direction	Expression
Look on the system	Neutral (EN)
(DON)	Smile (ES)
Look out the system	Odd face (EO)
(DOF)	Wry face (EW)

Table 4. Discrimination results(%)

State A	State B	Harm.	Total
52.5	65.1	58.2	62.2

Some symbolic labels were defined as an intermediate feature of the hierarchical discrimination. Both the acoustic and the visual events were labeled manually according to the occasion of the event. These labels were used as the supervisory signal for training the front-end neural network. Tables 2 and 3 show the labels of the acoustic and the visual events. In addition to the single-frame score, the temporal dynamics of the scores are also important for estimation of user's state. Therefore, we incorporated the first derivative of the scores into the feature set. The total number of dimensions of the intermediate features including differential coefficients was 26.

4 Experiment

4.1 Experimental Conditions

All neural networks were three-layer networks having input, hidden and output layers including the bias unit. We employed a softmax activation function at the output layer in order to obtain the outputs as the probability of the above-mentioned class. The activation function of the hidden layer was a logistic sigmoid function and the number of hidden units were determined by preliminary experiments. All experiments were conducted based on 5-fold cross validation. Here, we examined two-class discrimination because state C and the rest (i.e. state A and B) were separated clearly by latency.

4.2 Discrimination Results

The definitive results should be decided considering the time variation of the outputs since the scores of the user's state change frame by frame. In this paper, we decided the definitive class $\hat{c}$ as follows:

$$\hat{c} = \arg\max_{c}(\max_{1 \le t \le T}(p_{tc})) \tag{1}$$

where, T is the length of the segment for which the state is estimated and currently set to equal to the duration of each session. 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 1), therefore harmonic mean (denoted as Harm.) were employed for measuring the performance. Table 4 shows the definitive result when the best harmonic mean was obtained; the total accuracy was 62.2%. Some of the results of the back-end classifier were closed to our intention. For example, the score of state B tends to be high in the filler or self speech segments, and the score of state A tends to rise when the user moves his/her head in a short period. However, the performance is not enough to apply our method to actual dialog systems. One problem of the present method is the lack of an important feature for the estimation. The results of previous human evaluations have shown that gaze action is efficient for recognizing the state of the dialog partner. Additionally, we need to consider how to model the temporal structure of the user's behavior.

5 Conclusion

We investigated a method to estimate the user's internal state frame by frame during a dialog turn. Sequential features including MFCC, $\Delta F0$, zero cross ratio of speech and facial feature points were extracted continuously, and were used for the estimation. From the results of two-step discrimination, we obtained about 62% accuracy of the definitive results. However, it is necessary to improve the performance of the classification in order to apply the method to actual systems. To enhance the accuracy of the results, we will make two improvements: employ a feature representing eye movement and examine the model for modeling sequential data.

References

1. Forbes-Riley, K., Litman, D.: Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. Speech Communication 53(9), 1115–1136 (2011)
2. Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., Konos, H.: Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. Image and Vision Computing 27, 1760–1774 (2009)
3. Morrison, D., Wang, R., Silva, L.C.D.: Ensemble methods for spoken emotion recognition in call-centres. Speech Communication 49, 98–112 (2007)
4. Devillers, L., Vidrascu, L.: Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In: Proc. INTERSPEECH, pp. 801–804 (2006)
5. Chiba, Y., Ito, M., Ito, A.: Estimation of User's Internal State before the User's First Utterance Using Acoustic Features and Face Orientation. In: Proc. HSI (2012)
6. Saragih, J., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. Int. J. Computer Vision 91, 200–215 (2011)