# Find out What a User is Doing before the First Utterance: Discrimination of User's Internal State using Non-verbal Information

Yuya Chiba*, Seongjun Hahm* and Akinori Ito*
* Graduate School of Engineering, Tohoku University, Japan
E-mail: {yuya, branden65, aito}@spcom.ecei.tohoku.ac.jp

***Abstract*—In this research, we propose a method for estimating user's internal state (thinking or embarrassed) before the utterance toward a spoken dialogue system. Modeling user's internal state such as belief, skill or familiarity and introducing these model to the dialogue system should be useful to make flexible responses. However, because conventional estimation of internal state is based on the linguistic information of the previous utterance, it cannot estimate a user's internal state before the user's first utterance. We focus on a user's multimodal features such as filler word, silence, or face direction before the user's input utterance in order to model the user's internal state. The dialogue data were collected on the Wizard of Oz basis as training and test materials. Finally, we conducted an experiment for discrimination with two classification schemes and the hierarchical method obtained higher discrimination accuracy than that of pair-wise method.**

## I. Introduction

Speech-based man-machine interface such as a spoken dialogue system is expected to be a user-friendly interface because it can be used with the user's hands free and without training. However, spoken dialogue systems have been used only for limited task such as interactive voice response (IVR) of call centers or car navigation systems. One of reasons why spoken dialogue systems are not used in a wider situation is that these systems respond to the user's input in a uniform way, ignoring various intention of the user. To realize more flexible dialogue systems, many works have been done for introducing user models into dialogue systems [1]. Goal of these researches is to improve dialogue control by modeling a user's internal states. Here, the internal states represent various aspects of a user, such as belief [2], preference [3], skill [4,5], emotion [6] and familiarity to the system [7]. In these researches, a user is modeled to have several states and to change his/her response according to the current state. Then there are three major problems with user modeling for spoken dialogue systems: how to define internal states, how to estimate a user's current state and how to design dialogue system that exploits a user's current state. Among them, the state estimation problem is usually solved by observing dialogue history, especially the user's previous utterances. A problem of conventional estimation of internal state is that the estimation requires at least one utterance made by the user. When a user has trouble making the first utterance, the conventional method cannot estimate what kind of trouble the user has. This problem is especially serious for dialogue systems with small tasks that finishes with a user's one or two utterances. In such case, a conventional solution is to use heuristics such as incremental prompt [8].

As mentioned before, recognition of such troubles cannot be achieved using only a linguistic history of dialogue. Because the user's previous utterance is not available, we need to exploit not only audio information but also visual information for the recognition. Recognition of user's internal state using speech and facial image has been examined by a few researchs. For example, Gajsek et al. [6] attempted discrimination of speaker's emotion between rage and neutral, using MFCCs of the speech and the DCT coefficients of the facial image. Wöllmer et al. [9] proposed a method for discriminating normal and arousal emotions using speech features and feature points of a face. There have been a couple of systems that exploit estimation technique of user's emotion. Fujie et al. [10] developed a spoken dialogue system for a communication robot that considers user's positive and negative intention. Nomoto et al. [11] proposed a data mining system for call center recording based on emotion, especially anger of the speaker.

Our research aims to model user's internal state before a user makes the first utterance in order to determine whether the user have trouble making an utterance or not, and the cause of the trouble. In this research, we focus on two causes of a trouble speech interface users often face: one is that the user does not know what to speak, and the other one is that the user is taking time for preparing the utterance. These "states" have different aspect from user's belief or emotion treated in the previous works. The basic strategy for determining the user state is multimodal based on speech (filler words), silence, and visual feature such as face direction.

## II. Experimental Data

### A. Data collection

We collected dialogue video dips as training and test materials. There are two methodologies for collecting dialogue data: collecting acted dialogues using actors, and collecting natural dialogues using usual participants. Merit of the acted dialogue is easiness of collecting dialogues with various properties such as emotions and intentions. However, it is pointed out that an acted dialogue tends to be unnatural [12]. Therefore, we

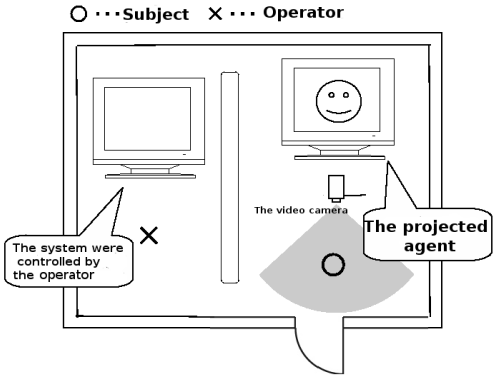


Fig. 1. Experimental circumstance

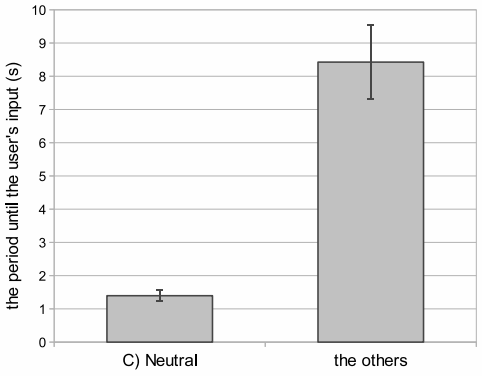


Fig. 2. The mean length of the period until user's input

collected real dialogues on a Wizard-of-Oz basis. The tasks of the dialogue are information retrieval task and "quiz" task. In the "quiz" task, the system asked a user a question and the user answers. Purpose of using the quiz task is to observe users' various reactions including "embarrassed" or "thinking." Nine subjects (eight males and one female) participated in the experiment. Fig.1 shows the experimental environment. Subjects were instructed to interact with an agent displayed on the monitor. Dialogue was actually controlled by the operator behind a partition. Subjects' utterances and frontal face were recorded using a digital video camera. One dialogue was a pair of system prompt utterance and the subject's answer utterance. We collected 199 dialogues, about 45 minutes in total.

### *B. Human Evaluation*

The collected dialogues were evaluated by five evaluators for labeling. They labeled a dialogue as one of the following three internal states: A) User was perplexed with the system's prompt utterance (embarrassment). B) User was thinking about the answer (thinking). C) Neither (Neutral). Table.I shows the results of the evaluation. The labels used in the later experiment were determined by majority vote by the five evaluators.

## III. THE DISCRIMINATION FEATURES

### *A. Speech-based features*

As mentioned before, we estimate the user's internal state without referring the user's previous utterance. Therefore, we should obtain features for estimating the internal state using audio signals observed from the beginning of the system's prompt utterance to the begining of the user's first utterance answering the prompt. Note that the user may make utterances other than the answer, such as filler words or interjections, which could be clues that indicate the user's confusion.

First, we examined the length between the end of the system's prompt utterance and the beginning of the user's answer utterance (denoted as $L_0$ hereafter) as a speech feature. From the evaluators' observation, the "neutral" dialogues tend to have shorter period between the end of the system prompt and the user's response. Here, the period contains silence, repairs and fillers. We manually determined this length for each dialogue. Fig.2 shows the mean length of this segment in the "neutral" and the other dialogues, where we can see large difference between the two types of dialogues.

Next, we investigated the audio signal between the beginning of the system's prompt to the end of the user's answer utterance in detail. As we can discriminate the state C and the other states using the feature explained above, the remaining problem is how to discriminate utterances of state A and B. To find features that contribute the discrimination, we classified the acoustic events in the observed signal into six classes shown in Table.II, then we investigated total length of events belonging to each class.

We investigated length of events of each class for all dialogues classified into state A or B, and observed difference of the length between the two internal states in order to find features useful for the discrimination. Let $N$ be the number of dialogues, $M_{ic}$ be the number of acoustic events of class $c$ observed in $i$-th dialogue, $L_{ic}$ be the total length of events belonging to class $c$ observed in $i$-th dialogue. Then we

TABLE I
THE EVALUATION RESULT

| Type | Agreement | Majority |
|---|---|---|
| A) | 14 | 20 |
| B) | 10 | 35 |
| C) | 81 | 140 |
| total | 105 | 195 |

TABLE II
THE CLASSIFICATION OF THE SPEECH SEGMENTS

| | |
|---|---|
| system | the segment of the system's utterance |
| user | the segment of the user's utterance |
| filler | filler of the user |
| repair | utterance modification by the user |
| etc | other user's voiced segment |
| breath | the aspirate or breath of the user |
| silence | the soundless segment |

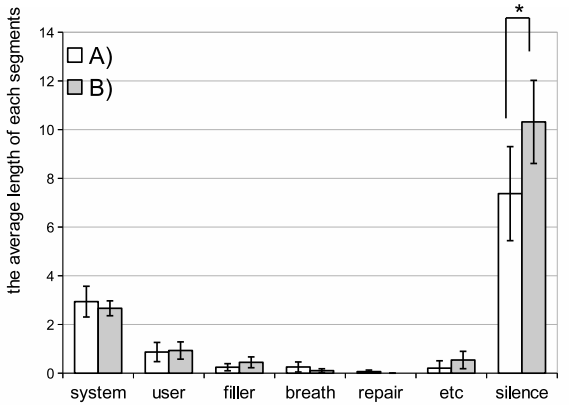


Fig. 3. $L_1$ of each class ($*p < 0.05$)

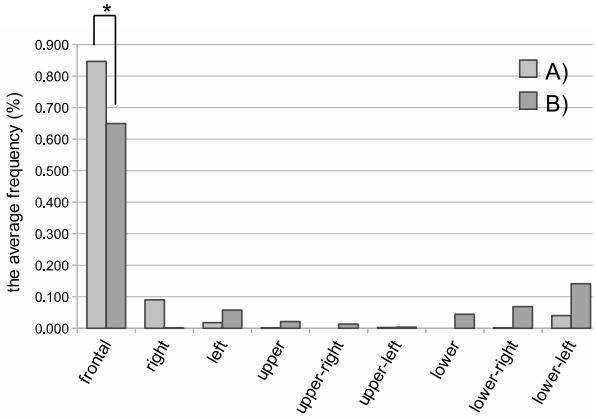


Fig. 5. The average frequency of the 9-oriented face orientation ($* p < 0.05$)

observed the length of events in a specific class in two aspects. The first one is length of the events normalized by number of dialogues:

$$L_1(c) = \frac{1}{N}\sum_{i=1}^{N} L_{ic} \tag{1}$$

and the other one is that normalized by number of events:

$$L_2(c) = \frac{\sum_{i=1}^{N} L_{ic}}{\sum_{i=1}^{N} M_{ic}} \tag{2}$$

$L_1$ and $L_2$ for each class are shown in Fig.3 and 4, respectively. We carried out unpaired t-test for each segment to seek the efficient features. Then we chose features that showed significant difference between the two classes at $5\%$ significance level. As a result, $L_1$ for fillers and $L_2$ for silences were chosen as the features. These facts indicate the subjects thinking the answer tend to be silent before giving an answer, and the long filler is considered to be sign of "thinking." According to these results, we chose the length of silence and filler as discrimination features.

### *B. Vision-based feature*

The visual features were selected in the same way as the speech features. We labeled face-orientation of the user during interaction with the system frame by frame using the recorded video. We manually labeled user's face-orientation as 9-oriented direction including "frontal". Fig.5 shows the distribution of face orientation. We conducted unpaired t-test on the frequency of face direction in state A and B. Then the significant difference is observed at the frequency of the "frontal" frames. Here, because we could not find significant difference between the left orientation and right orientation, we also examined using only vertical orientation of user's face. Fig.6 shows the distribution of the three face orientation. In this approach, the significant difference was obtained at the all orientation.

From these results, it is said that the users thinking about the answer are tend to turn their face from the system compared to the perplexed user, and face orientation is efficient for discriminating the user's internal state A and B.

## IV. THE DISCRIMINATION EXPERIMENT

For user's internal state discrimination, we designed feature vector by combining the four features, e.g. "the length after the system's prompt until the user's answer utterance", "the total length of filler segments", "the total length of silence segments" and "face orientation". Note that these features

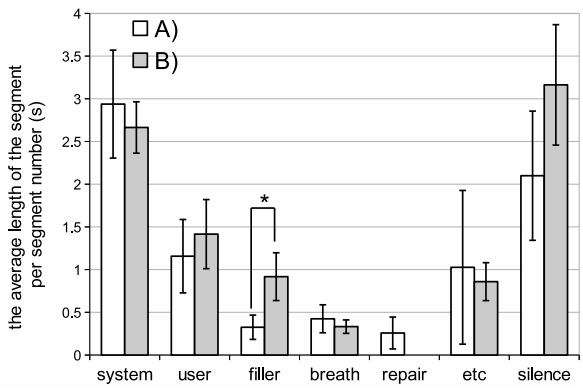


Fig. 4. $L_2$ of each class ($* p < 0.05$)

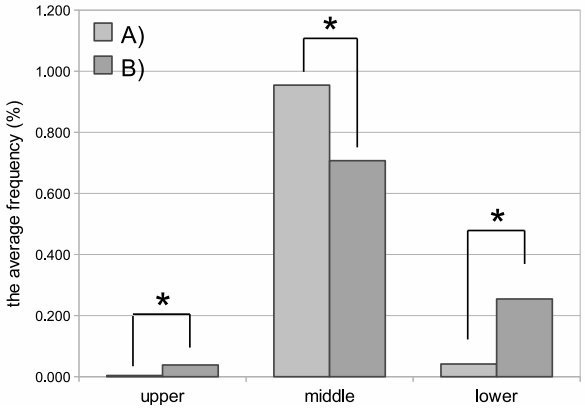


Fig. 6. The average frequency of the 3-oriented face orientation ($*p < 0.05$)

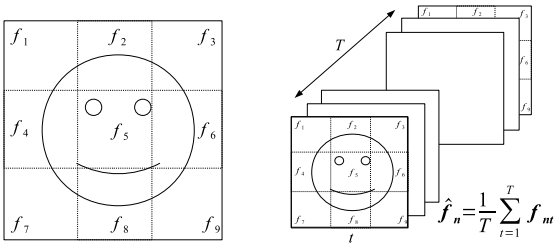


Fig. 7. The feature vector of face orientation

were based on manual labels. Automatic extraction of these features is an issue for future work.

### *A. the feature vector*

The values correspond to 9-oriented face orientation (see Fig.7) is expressed as a nine dimensional vector for the feature of each frame.

$$f_{nt} = \begin{cases} 1 & \text{if face orientation of frame } t \text{ is } n \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Then these features are averaged over the period. Let $T_1$ and $T_2$ be the frame when the system's prompt ends and the user's answer starts, respectively. We calculate the face orientation feature $\hat{f}_n$ as

$$\hat{f}_n = \frac{1}{T_2 - T_1} \sum_{t=T_1}^{T_2-1} f_{nt} \quad (4)$$

Next, we add the speech-based features to the face orientation features. Finally, the feature vector $\mathbf{v}$ is composed as follows.

$$\mathbf{v} = (\hat{f}_1, \hat{f}_2, \cdots, \hat{f}_9, L_0, L_{silence}, L_{filler}) \quad (5)$$

Here, $L_{silence}$ is the length of silence segment and $L_{filler}$ is the length of filler segment.

### *B. The experimental method and its result*

We carried out an experiment for discriminating the three classes of user's internal state using SVM. We examined two classification schemes: 3-class discrimination and hierarchical discrimination.

In the 3-class discrimination, we used libSVM with linear kernel. We used pair-wise method for multi-class discrimination. The experiment was carried out by cross-validation opened for each subjects.

In the hierarchical discrimination, we first discriminated dialogues into class C and others, and then "others" dialogues were classified into either "class A" or "class B."

The experimental results are shown in Table III. We obtained higher classification accuracy for dialogues of class C (normal), but the accuracy of the other two classes was not so high. The hierarchical classifier gave higher classification accuracy, especially for dialogues of class A (embarrassment).

TABLE III
THE EVALUATION RESULT(%)

| | Class A | Class B | Class C | Total |
|---|---|---|---|---|
| 3-class | 25.0 | 51.4 | 95.7 | 83.1 |
| Hierarchical | 40.0 | 65.8 | 95.0 | 84.1 |

## V. CONCLUSION

In this paper, we proposed a method for estimating the internal state of a user of a spoken dialogue system for the user's first utterance. In addition to "normal" state, we assumed two internal states: state A (the user is perplexed by the system's utterance) and B (the user is thinking how to answer the system). From the experiment, we found three useful features based on speech: "The period until user's input", "the period of the filler segment" and "the period of the silence segment", as well as vision-based feature: "face orientation." We conducted an experiment for estimating user's internal state, and obtained higher accuracy using the hierarchical classification method. As a future work, we will investigate other features from multimodal responses of the user and examine dynamic discrimination method.

## REFERENCES

[1] R. Kass and T. Finin, "Modeling the user in natural language systems", Computational Linguistics, Vol 14, No.3, 1988
[2] A. Kobsa, "User Modeling in Dialog Systems: Potentials and Hazards", AI&Society, vol. 4, pp. 214-140, 1990.
[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, pp. 329-351, 2004.
[4] K. Komatani, S. Ueno, T. Kawahara and H. G.Okuno, "Flexible guidance generation using user model in spoken dialogue systems", Proc. COLING, 2003.
[5] K. Jokinen and K. Kanto, "User Expertise Modelling and Adaptivity in a Speech-based E-mail System", Proc. COLING, 2004.
[6] R. Gajsek, V. Struc, S. Dobrisek and F. Mihelic," Emotion Recognition using Linear Transformations in Combination with Video ", Proc. Interspeech, pp1967-1970, 2009.
[7] F. de Rosis, N. Novielli, V. Carofiglio, A. Cavalluzzi and B. De Carolis, "User modeling and adaptation in health promotion dialogs with an animated character", J. Biomedical Informatics, vol. 39, pp.514-531, 2006.
[8] N. Yankelovich, "How do users what to say?", Interactions, vol. 3, no. 6, pp. 32-43, 1996.
[9] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller and S. S. Narayanan,"Context-Sensitive Multimodal Emotion Recognition from Speech and Facial Expression using Bidirectional LSTM Modeling", Proc. Interspeech, pp. 2362-2365, 2010.
[10] S. Fujie, Y. Ejiri, H. Kikuchi, and T. Kobayashi, "Recognition of positive/negative attitude and its application to a spoken dialogue system", System and Computers in Japan, Volume 37, Issue 12, pp45-55, 2006
[11] N. Nomoto, H. Masataki, O. Yoshioka and S. Takahashi, "Detection of Anger Emotion in Dialog Speech Using Prosody Feature and Temporal Relation of Utterances", Proc. Interspeech, pp. 494-497, 2010.
[12] A.Batliner, K. Fischer, R. Huber, J. Spilker and E. Nöth, "Desperately seeking emotions or: actors, wizards, and human beings.", Proc. ISCA Workshop on Speech and Emotion, pp. 195-200, 2000.
[13] C. T. Ishi, H. Ishiguro and N. Hagita, "Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality", Speech Communication, vol. 50, pp. 531-543, 2008.
[14] K. Komatani, N. Kanda, T. Ogata and H. G. Okuno, "Contextual Constraints based on Dialogue Models in Database Search Task for Spoken Dialogue Systems", pp877-880, ISCA, 2005.