Hindawi Publishing Corporation
Advances in Human-Computer Interaction
Volume 2012, Article ID 865362, 10 pages
doi:10.1155/2012/865362

Research Article

Estimating a User's Internal State before the First Input Utterance

Yuya Chiba and Akinori Ito

Graduate School of Engineering, Tohoku University, 6-6-5 Aramaki aza Aoba, Aoba-ku, Sendai, Miyagi 980-8579, Japan

Correspondence should be addressed to Yuya Chiba, yuya@spcom.ecei.tohoku.ac.jp

Received 16 February 2012; Revised 30 April 2012; Accepted 4 May 2012

Academic Editor: Kerstin S. Eklundh

This paper describes a method for estimating the internal state of a user of a spoken dialog system before his/her first input utterance. When actually using a dialog-based system, the user is often perplexed by the prompt. A typical system provides more detailed information to a user who is taking time to make an input utterance, but such assistance is nuisance if the user is merely considering how to answer the prompt. To respond appropriately, the spoken dialog system should be able to consider the user's internal state before the user's input. Conventional studies on user modeling have focused on the linguistic information of the utterance for estimating the user's internal state, but this approach cannot estimate the user's state until the end of the user's first utterance. Therefore, we focused on the user's nonverbal output such as fillers, silence, or head-moving until the beginning of the input utterance. The experimental data was collected on a Wizard of Oz basis, and the labels were decided by five evaluators. Finally, we conducted a discrimination experiment with the trained user model using combined features. As a three-class discrimination result, we obtained about 85% accuracy in an open test.

1. Introduction

Speech is the most basic medium of human-human communication and is expected to be one of the main modalities of more flexible man-machine interaction along with various intuitive interfaces rather than traditional text-based interfaces. One major topic of speech-based interfaces is the spoken dialog system. Studies on spoken dialog systems have attempted to introduce a user model, which models the user's internal states, to make the dialog more flexible. The user's internal states represent various aspects of the user, such as belief [1], preference [2], emotion [3], and familiarity with the system [4–6]. These aspects can also be categorized according to their persistency: as the user's knowledge and preference are persistent, they can be used for personalizing of the dialog system [7]. Other internal states such as emotion or belief are transient and so are used for making a dialog more natural and smooth. These kinds of internal state should be estimated session-by-session. In this paper, we focus on the latter, transient states.

These internal states are estimated based on the verbal and nonverbal information included in the interaction between the user and the system. In this work, we consider a system-initiative dialog system that presents a prompt message at the beginning of a session. In such a system, a session between the user and the dialog system can be divided into three phases: before the system prompt (*Phase 1*), after the prompt and before the user's response (*Phase 2*), and the rest (*Phase 3*). Figure 1 shows the three phases in a session.

Many conventional studies on user modeling have focused on the linguistic information of the user's utterance and estimated the user's internal states based on the dialog history (i.e., previously observed utterances made by the user and the system) [8–10]. As these works use the user's utterances as the basis for estimating the internal states, these methods can only be used in *Phase 3*, because no linguistic information (history of user utterances) is available in *Phase 1* or *Phase 2*.

Recognition of the internal states of the conversation partner (i.e., the user of the system) is related to social interaction in human-human communication. Studies on the social interaction of human-computer interfaces have included conversations with robots [11–13] and virtual agents [14–17]. An important cue to recognize social interaction is nonverbal information. These studies employ some multimodal information such as hand gestures, head nods, face direction, and gaze direction as well as spoken language

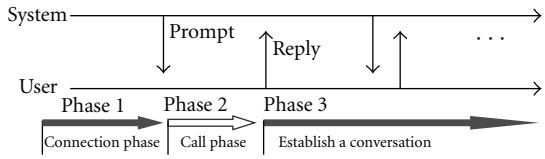

Figure 1: Phases of the dialog session.

to build teamwork in the collaboration with a robot [11, 12] or to create a chance to address the user [13]. Meanwhile, Maatman et al. [14] and Kopp et al. [15] studied the natural behavior of the agent while the user is speaking. These virtual agents need to generate nonverbal outputs to the conversation partner, and these outputs directly affect the naturalness of the dialog. Buß and Schlangen [18] defined the short utterance segment in continuous speech as a subutterance phenomenon and analyzed the roles of the turn-taking or back channels. These works also focused on *Phase 3* of a session.

On the other hand, several works investigated the internal states in *Phase 1*. Hudson et al. [19] and Begole et al. [20] discriminated a user's *interruptibility* by analyzing subjects in an office and a user of instant messages, respectively. In addition, Satake et al. [21] studied the appropriateness of the approaching behavior when addressing people in human-robot dialog.

In the present work, we investigate how to estimate the user's internal state in *Phase 2*. We believe that user modeling in *Phase 2* is important, even though it has not been investigated to date. First we explain the issue of user modeling in *Phase 2* in detail. For example, if the user does not understand the meaning of the system's prompt, he or she could abandon the session without uttering a word (case 1). The other possibility is that the user is considering how to answer the prompt (case 2). In both of these cases, the conventional systems respond uniformly even though different responses are desired, because they treat both cases as "the user did not know what to say."

In real systems, a heuristic solution such as *incremental prompt* [22] is employed, where the system offers a different prompt if the user does not respond within a certain time after the first prompt. The system should provide more detailed information to clarify the intention of the prompt in case 1; however, in case 2, intervention from the system in the dialog is undesirable. In fact, Kobayashi et al. [17] described that a system response that does not consider the difference between cases 1 and 2 will confuse the user. This problem is especially serious for systems with dialogs that finish in one or two user utterances.

We considered these two dialog cases and assumed that they are derived from different internal states. Here, we define three internal states. In the first one (State A), the user does not know how to answer the prompt. In the second one (State B), the user is taking time to consider the answer. In the third one (State C), the user has no difficulty in answering the system.

The goal of our task is to discriminate the internal state of a user among these three States A, B, and C by observing the user's behavior. This challenge is close to the research on turn-taking of the dialogue. There have been a number of works that analyze turn-taking behavior for human-human dialogs [23] as well as human-machine dialogs [24–28]. Introducing the turn-taking mechanism to a spoken dialog system is believed to make the dialog system to interact to the user in more natural and effective way. Notably, Edlund and Nordstrand [29] have researched on the multimodal dialog system and examined the turn-taking between the user and an agent (an animated talking head) that made a gesture such as head motion and gaze control.

Most of these works focused on the turn-taking behavior after the dialogue establishment (i.e., *Phase 3*). Some of the knowledge obtained from these works might be useful for our problem; however, it seems to be difficult to discriminate the user's internal state at *Phase 2* such as State A and State B using only cues for turn-taking.

To discriminate the user's internal state without observing the user's verbal utterances, we exploit audio and visual features for the estimation and investigate which features can be used for discriminating these internal states. Ideally, all features should be extracted automatically and the discrimination should be made incrementally; however, in this paper, we manually labeled the data to extract a part of the features and used the entire video sequence for discrimination, because the objective of this paper is to investigate which feature can be used for this purpose. Automatic extraction of the useful features and incremental discrimination are issues for future works.

This paper is organized as follows. Preparation of the experimental data is described in Section 2. Then the audio features and visual features are introduced in Sections 3 and 4, respectively. Finally, the results of the experiments are presented in Section 5.

2. Collection and Analysis of Dialog Data

2.1. Dialog Tasks. We collected dialog data to analyze internal states of the users. There are two possibilities for collecting dialog data: collecting acted dialogs using actors and collecting natural dialogs using naïve participants. The merit of acted dialogs is ease of collecting dialogs with various properties such as emotions and intentions, but such dialog tends to be unnatural [30]. Therefore, we decided to collect natural dialogs.

For the experiments, we prepared two tasks: (1) a simple information retrieval task and (2) a "question and answer" task. The information retrieval task simulated a restaurant guidance system. This system was task-oriented, and the user has to answer to a series of prompts to achieve the task. For example,

> System: Are you looking for a restaurant of any region?
>
> User: At region A.
>
> System: How much is your budget?
>
> ⋮

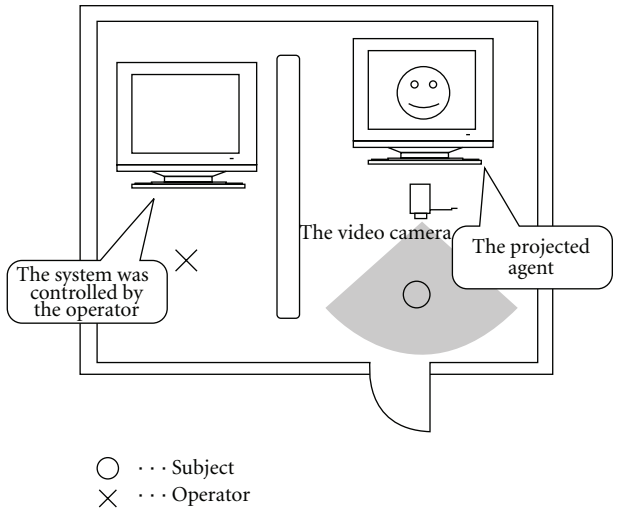

Figure 2: The experimental environment.

In the experiment, the operator required the subjects to achieve ambiguous task to search an "affordable" restaurant of Japanese food.

Because we thought the user's internal states are not expressed frequently by using a natural system, we prepared the "question-and-answer" task, where the system asks a question and the user answers, in order to make the user express their internal states. An example of the "question-and-answer" session is as follows:

System: What is the date today?

User: Uhm. . ., it's May . . . May 17th today.

The questions were independent of each other and presented at random.

2.2. Data Collection Procedure. The data collection was carried out on a Wizard-of-Oz basis. We prepared an agent with a simple cartoon-like face and synthesized voice and displayed it on an LCD monitor to encourage the user to pay attention to the front of the system. We prepared several faces with emotion (neutral, joyful, angry, sad) for the agent. The agent had always the neutral face at the beginning of the system prompt and changed its facial expression after the user's utterance. The expression was decided by the operator (the "wizard") according to appropriateness of the user's response.

Figure 2 shows the experimental environment. We placed a digital video camera between the monitor and the user and recorded the user's frontal face. The operator (the first author of this paper) operated the system behind the partition. In either task, the prompt was repeated at 15-second intervals if user did not make the input utterance.

After recording a dialog, we segmented the recorded dialog into "sessions," where one session included the system's prompt and the user's response. When a user's utterance was not observed, that part (from the beginning of the system's prompt to the beginning of the next prompt) was used as a session.

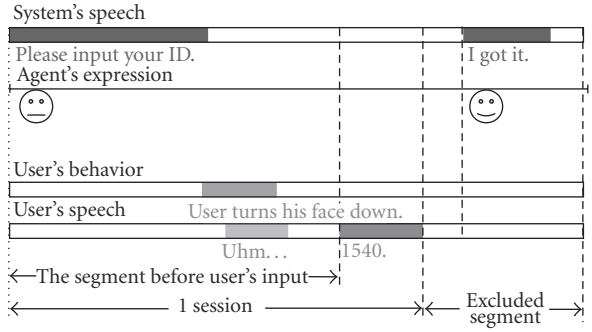

Figure 3: Outline of dialog data.

Figure 3 shows an overview of one dialog session. We excluded the segment after the beginning of the user's input utterance from a session because our interest is the duration before the user's utterance. Because the agent's facial expression changes after the user's utterance, the influence of facial expression change of the agent to the user's attitude was also excluded from the later analysis.

As we split one dialog into more than one session, some of the sessions were turns in *Phase 3*. Here, the user's internal state at *Phase 3* just after the system's prompt is supposed to be quite similar to that at the *Phase 2*, except for an existence of contextual information. Due to this, we used all the "sessions" of the dialog data for the later experiment even if the session belonged to *Phase 3*. The estimation of the user's internal state at *Phase 3* is expected to be improved by using the contextual information; however, this issue is not covered here.

We asked nine subjects (eight males and one female) to make conversation with the dialog system. Number of dialogs for one subject was different from subject to subject. The total number of session was 199 (22.1 sessions per user in average, $\sigma^2 = 3.43$).

2.3. Subjective Evaluation of the Internal States. Next, we evaluated the users' internal states to make "ground truths" of the internal states. There could be two possibilities for making ground truths: one is based on the user's introspection or intuition, and the other one is based on the observer's opinion. Here, the goal of our work is to develop a multimodal dialog system that behaves like a human receptionist, who should determine the user's internal state by only observing the user's behavior. Therefore, we decided to evaluate the sessions by evaluators' observation. Agreement of these two evaluations is an interesting issue [31] to investigate in a future work.

The gathered sessions were labeled by five evaluators. They were asked to watch the recorded video of the entire session and classify each session into one of three states:

State A: The user was perplexed by the system's prompt

State B: The user was considering an answer to the prompt

State C: Neither of the above (neutral)

Table 1: Evaluation results (Agreement).

	Task (1)	Task (2)	Total
State A	1	9	10
State B	0	14	14
State C	48	33	81
Total	49	56	105

Table 2: Evaluation results (Majority).

	Task (1)	Task (2)	Total
State A	3	17	20
State B	2	33	35
State C	69	71	140
Total	74	121	195

Table 3: Concordance ratio of each evaluator's result.

	E1	E2	E3	E4	E5
E1	—	.74	.71	.73	.71
E2	—	—	.90	.89	.65
E3	—	—	—	.91	.65
E4	—	—	—	—	.68
E5	—	—	—	—	—

Table 1 shows the number of matches of evaluations by five evaluators and Table 2 shows the results of a majority vote. As we expected, user's internal state such as State A and B were often appeared in the task (2). Moreover, concordance ratio of each evaluator's decision is shown in Table 3. The concordance ratio was calculated by the following equation:

$$\text{Conc.} = \frac{n_{ij}}{N}. \quad (1)$$

Here, n_{ij} is the number of matched label between the evaluator i and the evaluator j, and N is the total number of the sessions ($N = 199$). The results show that the decision of the evaluator E2, E3, and E4 are well accorded; however, the decision of evaluator E5 does not match with that of others too much. Therefore, we decided to use 195 sessions by the majority vote as the experimental data. Four sessions were excluded because they had the same number of votes for two different labels.

As mentioned, we estimate the user's internal state without referring to the user's previous utterance (i.e., *Phase 2*). Therefore, we have to obtain features for estimation from audio signals observed in the segment before the user's input. Note that the user may make utterances other than the answer, such as filler words or interjections, which could provide clues as to the user's internal state.

3. Speech-Based Features

3.1. The Length until User's Input. First, we examined the length between the end of the system's prompt utterance and the beginning of the user's answering utterance (denoted as L_0 hereafter) as a speech feature. This period contains silence, repairs, fillers, and breathy voice of the user. We manually determined this length for each dialog. Figure 4 shows the mean length of this segment in the "neutral" and "other" dialogs, where we can see large differences between the two types of dialog.

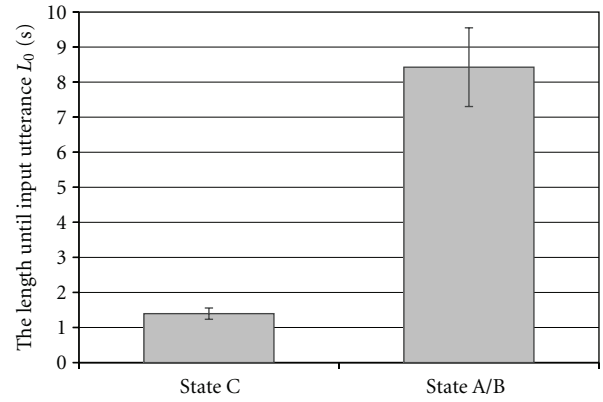

Figure 4: Mean length of the period until user's input.

Table 4: Classification of speech segments.

System	System's prompt utterance
User	Input utterance
Filler	Filler utterance
Repair	Repair utterance
Etc	Other voiced segment
breath	Aspiration or breath
silence	Soundless segment

This result reflects the fact that the evaluators tended to label the user's internal state as "neutral" in sessions where the user answered to the prompt immediately.

3.2. The Length of Speech Classification. Next, we investigated the audio signal of the segment before the user's input in detail. As we can discriminate State C and the other states using the feature explained above, the remaining problem is how to discriminate dialog of States A and B.

To find features that will assist the discrimination, we classified the acoustic events in the observed signal into six classes as shown in Table 4, then investigated the total length of events belonging to each class. Among the classes shown in Table 4, the "system" segment is for utterances by the system, and all of the other classes are for utterances by the user.

We investigated the length of events of each class for all dialogs classified into States A and B and observed the difference in length between the two internal states in order to find features effective for discrimination. Let N be the number of dialogs, M_{ic} the number of acoustic events of class c observed in the ith dialog, and L_{ic} the total length of events belonging to class c observed in the ith dialog. Then we observed the length of events of a specific class using two

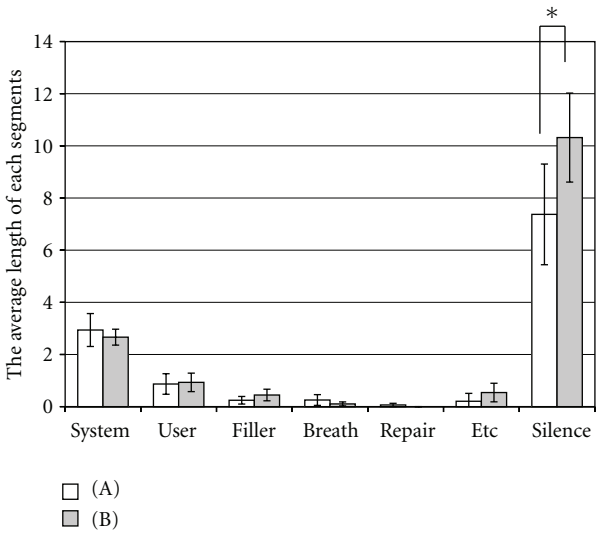

Figure 5: L_1 of each class ($^*P < 0.05$).

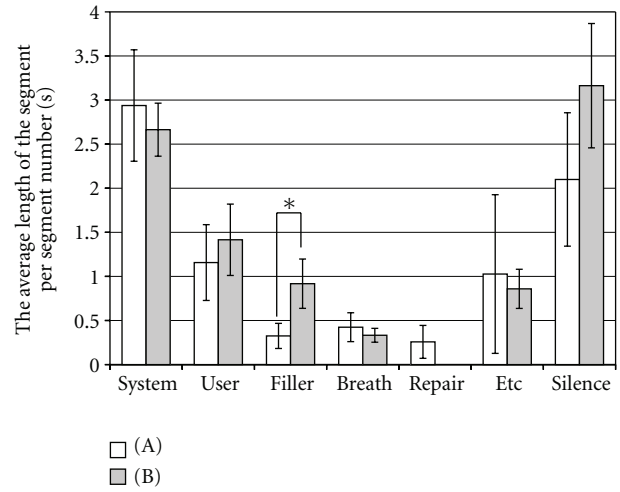

Figure 6: L_2 of each class ($^*P < 0.05$).

normalization methods. The first one is the length of events normalized by number of dialogs:

$$L_1(c) = \frac{1}{N}\sum_{i=1}^{N} L_{ic} \tag{2}$$

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

$$L_2(c) = \frac{\sum_{i=1}^{N} L_{ic}}{\sum_{i=1}^{N} M_{ic}}. \tag{3}$$

The results yielded by the two normalization methods are different because of the difference in frequency of events. Using L_1, the value for a rare event tends to be small, whereas the value can be larger when evaluating it by L_2.

L_1 and L_2 for each class are shown in Figures 5 and 6, respectively. We performed the unpaired t-test for each segment to decide the efficient features, then chose those features that showed a significant difference between the two classes at the 5% significance level. As a result, significance differences are observed at L_1 for fillers and L_2 for silences. Therefore, we chose the length of the filler and silence as efficient features for discriminating between states A and B. These facts indicate that subjects who were thinking about the answer (e.g., who were labeled as state B) tended to be silent before answering, and so a long filler was considered to be a sign of "thinking."

3.3. Length of Filled Pause. Upon analyzing the sessions of state B, we found that vowels tended to be lengthened in utterances other than filler words. This phenomenon is called *filled pause*. Goto et al. proposed a method to detect filled pauses [32] and described that filled pauses serve to maintain the speaker's turn and to express the mental state while thinking of the next utterance.

Goto et al. focused on the features of filled pauses such as little F0 and spectrum variation and formulated the filled pause likelihood. Here, the value of the F0 transition and the spectral envelope deformation at frame t are defined as $A_f(t)$ and $A_s(t)$. $A_f(t)$ is obtained as the slope of linearized temporal transition of F0. Meanwhile, $A_s(t)$ is the product of the slope and the fitting error of the temporal transition of the linearized spectral envelope. The F0 transition and spectral envelope transition are linearized by least-squares fitting. Then, the filled pause likelihood is calculated as

$$P_{fp}(t) = \exp\left(-\frac{\left(RS_f(t) + (1-R)S_s(t)\right)^2}{W^2}\right). \tag{4}$$

Here, $A_s(t)$ and $A_f(t)$ are averaged by a short period, that is,

$$S_i(t) = \frac{1}{\text{Period}_{fp}} \sum_{\tau=0}^{\text{Period}_{fp}-1} A_i(t-\tau) \quad (i = f, s). \tag{5}$$

Period_{fp} is 10 frame shifts. R and W are heuristic values, and we set them as $R(0.011)$ and $W(1.0)$ here. Finally, Goto et al. calculated the accumulated sum of $P_{fp}(t)$ as long as $P_{fp}(t) > e^{-1}$. If the sum of P_{fp} at frame t is larger than a certain threshold (we set this value as $9.5e^{-1}$), then the segment of frame t is judged to be a filled pause segment.

Filled pause features are thought to affect the choice of label by the evaluators. We extracted the filled pauses contained in the user's utterance using Goto's method and used the length of filled pause (L_{fp}) as a feature for discriminating States A and B.

4. Vision-Based Features

4.1. Distribution of Face Orientation. We analyzed the face orientation of the user in the segment before the user's input as a visual feature. First, we investigated the tendency of the face orientation in the dialog data. We manually labeled the user's face orientation as one of nine directions including frontal. Figure 7 shows the distribution of face orientation. We conducted the unpaired t-test on the frequency of face

Table 5: Feature points of the face.

	A	B
Yaw	Center of the nose region	Center of the face region
Roll	Center of the left eye	Center of the right eye
Pitch	Center of the eyes and the nose	Center of the face region

direction in States A and B and found a significant difference in the frequency of the frontal frames. This result shows that the users who were considering the answer tended to turn their face away from the system compared to the perplexed users. From this observation, face orientation is considered to be efficient for discriminating a user's internal states A and B.

4.2. Face Orientation Feature. As mentioned above, the face orientation feature is thought to be effective for identifying a user's internal state, but the classification into nine orientations is too coarse and was too difficult for the actual system to use because the labels were assigned manually. Therefore, we carried out automatic estimation of face orientation and estimated the face direction as a continuous value.

First, we detected the parts of the face (eyes and nose) in each frame. This was done by template matching of the face region in a frame, which was determined through face detection [33] and tracking [34]. We then checked the detection results and corrected all the misdetected frames manually to investigate the efficiency of the feature excluding the effect of estimation error.

Next, we calculated three-dimensional face orientation (*yaw*, *roll*, and *pitch*) based on the face region and positions of the parts. We approximated the shape of the human head using a sphere and used a sine value instead of the angle (Figure 8)

$$\begin{aligned} yaw &\equiv \frac{2x}{r} = \frac{2(A_x - B_x)}{r}, \\ roll &\equiv \frac{b}{a} = \frac{B_y - A_y}{\sqrt{(B_x - A_x)^2 + \left(B_y - A_y\right)^2}}, \\ pitch &\equiv \frac{2y}{r} = \frac{2\left(A_y - B_y\right)}{r}. \end{aligned} \tag{6}$$

Here, r is the diameter of the head, estimated as the width of the face region. The points $A = (A_x, A_y)$ and $B = (B_x, B_y)$ are feature points in a frame, as shown in Table 5. We calculate *yaw*, *roll*, and *pitch* frame by frame and denote these values at frame t as $yaw(t)$, $roll(t)$, and $pitch(t)$, respectively.

Figure 9 shows an example of the face orientation calculation. In this example, the user turns his face from the front to the lower right from frames 50 to 80. There is a large change of *yaw* and *pitch* around frames 50 to 80, as well as a small change of *roll*.

4.3. Data Compression. The results of a preliminary examination showed that the simple descriptive statistics of face orientation (e.g., mean, variance, maximum, and minimum) were not effective for discrimination. Therefore, we tried to treat the face orientation feature as sequential data. However, the number of values of face orientation depends on the number of frames of the session and varies from around 100 to 700. To simplify the calculation of discrimination function, we need to extract feature vectors of fixed dimension from these face orientation values. To achieve this, we used the piecewise aggregate approximation (PAA) method [35], which linearly compresses the face orientation vectors into a fixed number of vectors.

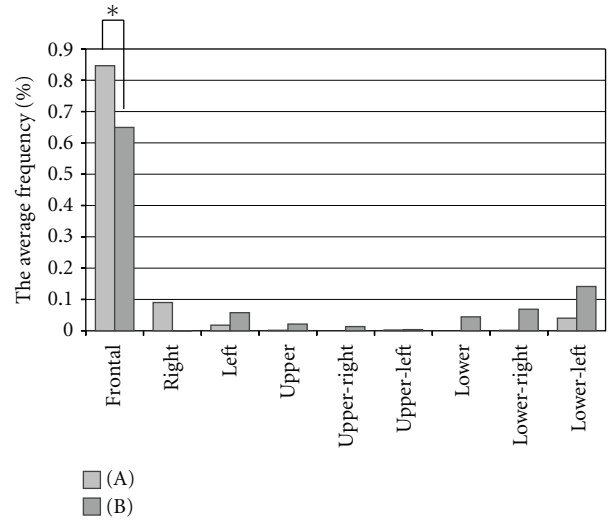

Figure 7: Average frequency of the 9-oriented face orientation ($^*P < 0.05$).

The compressed feature vectors $\overline{x}_1, \ldots, \overline{x}_n$ are calculated from the face orientation vectors $x_1, \ldots, x_N$ $(n \leq N)$ as follows:

$$\overline{x_i} = \frac{n}{N} \sum_{j=(N/n)(i-1)+1}^{(N/n)i} x_j. \tag{7}$$

Figure 10 shows an example of a 110 point face orientation sequence compressed into 15 dimensions by PAA.

The original data of this example is the same as the uppermost one in Figure 9, and we found that the number of data is reduced while retaining the rough deviation of the original signal.

5. Discrimination Experiment

5.1. Experimental Method. We carried out an experiment for discriminating the three classes of the user's internal state (e.g., State A, B, and C) using the Support Vector Machine (SVM) [36]. The elements of the feature vector are employed from the features examined above (details of the features will be described in Section 5.2). We used libSVM [37] with a linear kernel for the experiments. The simple pairwise method was employed for multiclass discrimination. The experiments were carried out by cross-validation opened for each subject. Therefore, we conducted a 9-fold validation

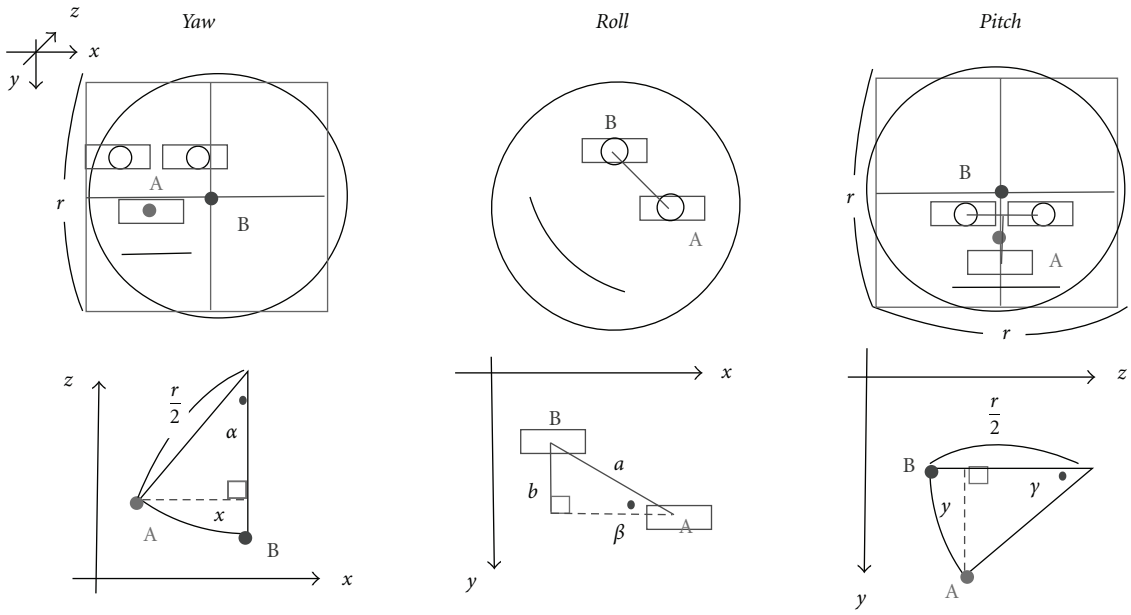

Figure 8: Calculation of face orientation feature.

Table 6: Elements of feature vectors.

Speech-based feature		(a)	(b)	(c)	Face orientation feature		(a)	(b)	(c)
Length until input utterance	L_0	○	○	○	continuous ($t = 1, \ldots, 15$)	$\overline{yaw}(t)$		○	○
Length of silence	L_{silence}	○	○	○		$\overline{roll}(t)$		○	○
Length of filler	L_{filler}	○	○	○		$\overline{pitch}(t)$		○	○
Length of filled pause	L_{fp}			○	discrete ($n = 1, \ldots, 9$)	$\hat{f}_n$	○		

Table 7: Discrimination result.

	(A)	(B)	(C)	Total
Set (a)	15.0	68.6	97.1	83.6
Set (b)	45.0	62.9	96.4	85.1
Set (c)	50.0	62.9	96.4	85.6

test, in which the amounts of training data and test data of each fold were unequal.

5.2. Feature Set. We prepared two feature sets for the discrimination because we have investigated both "discrete" and "continuous" face orientation features. The discrete face orientation is represented by face direction symbols decided manually. We defined the feature set including the discrete face orientation feature as set (a). On the other hand, the continuous face orientation feature is calculated by the previously explained image processing and compressed by PAA. The feature set including the continuous face orientation feature is defined as set (b). Furthermore, we prepared the feature set (c) to examine the effect of the length of filled pause.

Feature Set (a). The values correspond to the nine face orientations (see Figure 11) expressed as a nine-dimensional vector for the feature of each frame

$$f_{nt} = \begin{cases} 1 & \text{if face orientation of frame } t \text{ is } n, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

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

$$\hat{f}_n = \frac{1}{T_2 - T_1} \sum_{t=T_1}^{T_2-1} f_{nt}. \tag{9}$$

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

$$\mathbf{v} = \left(\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_9, L_0, L_{\text{silence}}, L_{\text{filler}}\right). \tag{10}$$

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

Feature Set (b). Feature set (b) includes the continuous face orientation value as described in Section 4.2. This set was

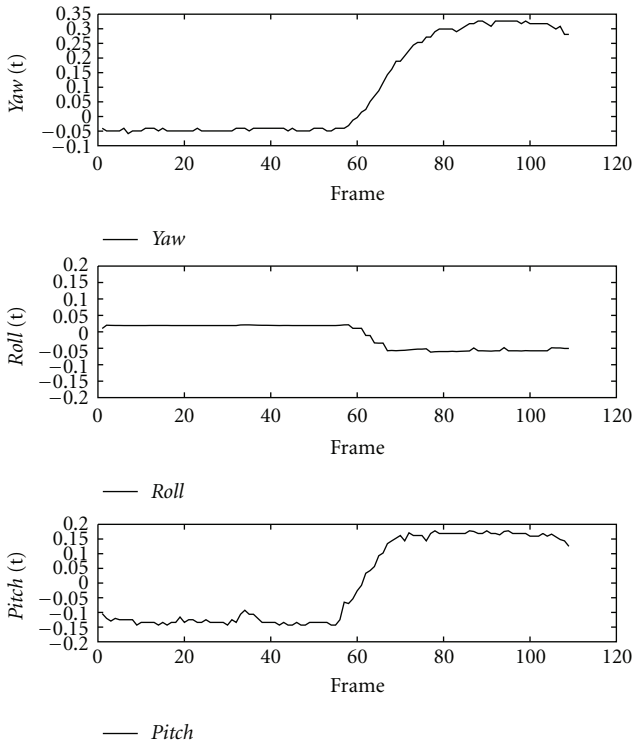

Figure 9: Example of sequential face orientation data when the user turns his face from frontal to lower right.

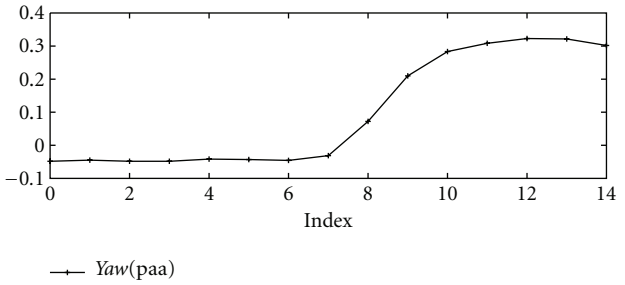

Figure 10: Example of frame number reduction.

prepared to compare the effectiveness on the discrimination between the continuous face orientation feature and the discrete one. We employed compressed face orientation data. According to a preliminary experiment, we decided to compress the face orientation vectors into the 15 points ($n = 15$) that gave the best classification accuracy. This set also included three speech-based features that were same as those in the feature set (a), and the total number of dimensions of feature vector **v** was 48.

Feature Set (c). This feature set was prepared to examine the effect of the filled pause feature. In addition to the feature in set (b), the feature set (c) included the length of filled pause L_{fp}. Therefore, the total number of dimensions of feature vector **v** was 49. The elements of each feature vector are summarized in Table 6.

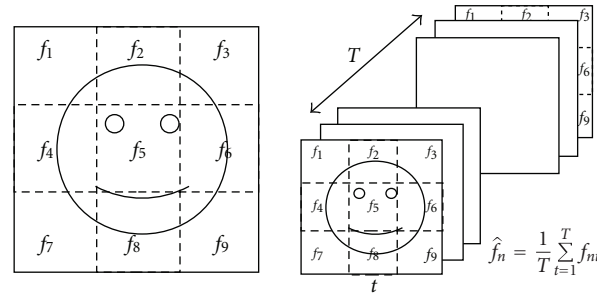

Figure 11: The feature vector of face orientation.

5.3. Experimental Results. Finally, we carried out an experiment using the features using the above-mentioned feature sets. The results are shown in Table 7. Each row of Table 7 shows the discrimination results for each feature set. From Table 7, we find that the discrimination accuracy using feature set (b) was higher than feature set (a), and that using feature set (c) was the highest. To validate the statistical significance of the difference among the total discrimination rates, we conducted one-way repeated measures ANOVA, where the feature set was the factor and nine results obtained by different cross validation experiments were the repeated measure. As a result, we could not find any significant difference among the three feature sets.

Although we could not observe significant differences among the features, we can conclude that the continuous face orientation feature is effective compared with the face orientation distribution because the continuous face orientation can be obtained automatically, which is indispensable for realization of automatic estimation of the internal state. The filled-pause feature did not give significant improvement either, but there might be a possibility to obtain more improvement by combining the filled-pause feature with other multimodal features.

6. Conclusion

In this paper, we defined three internal states (State A, B, and C) of a user of a dialog system before the first user utterance and investigated methods of modeling these states. It is important to estimate the user's internal state before the user's input in order to make the response of the system more appropriate, but this issue has not been studied to date. In this paper, we focused on the speech-based and face orientation features before the user's input because they are considered to express the user's internal state. As speech-based features, we used four features: the length until the user's input, the length of filler segments, the length of silent segments, and the length of filled pauses. As face orientation features, we used the sequence of three-dimensional face rotation angles and discrete face orientation frequencies. From the results of discrimination experiments, we examined the efficiency of these features. We obtained the discrimination accuracy as high as 85.6%, but we could not observe significant differences between the two face orientation features.

A remaining problem is that, the features proposed in this study are not available until the user's input utterance is observed, because we examined the segment until "just" before the user's input. In addition, the speech-based features and face orientation features were examined independently; however, the correlation between both features is important in practice. Therefore, in a future work, we will employ a sequential discrimination method such as HMM or CRF to analyze and select the features. Moreover, the total number of data in our experiment was not enough, so we need to collect more data and confirm the validity of our proposed discrimination method and features.

References

[1] A. Kobsa, "User modeling in dialog systems: potentials and hazards," *AI & Society*, vol. 4, no. 3, pp. 214–231, 1990.

[2] A. N. Pargellis, H. K. J. Kuo, and C. H. Lee, "An automatic dialogue generation platform for personalized dialogue applications," *Speech Communication*, vol. 42, no. 3-4, pp. 329–351, 2004.

[3] R. Gajšek, V. Štruc, S. Dobrišek, and F. Mihelič, "Emotion recognition using linear transformations in combination with video," in *Proceedings of the Interspeech*, pp. 1967–1970, 2009.

[4] K. Jokinen, "Adaptation and user expertise modelling in AthosMail," *Universal Access in the Information Society*, vol. 4, no. 4, pp. 374–392, 2004.

[5] F. D. Rosis, N. Novielli, V. Carofiglio, A. Cavalluzzi, and B. D. Carolis, "User modeling and adaptation in health promotion dialogs with an animated character," *Journal of Biomedical Informatics*, vol. 39, no. 5, pp. 514–531, 2006.

[6] K. Komatani, S. Ueno, T. Kawahara, and H. G. Okuno, "Flexible guidance generation using user model in spoken dialogue systems," in *Proceedings of the COLING*, pp. 256–263, 2003.

[7] C. A. Thompson, M. H. Göker, and P. Langley, "A personalized system for conversational recommendations," *Journal of Artificial Intelligence Research*, vol. 21, pp. 393–428, 2004.

[8] S. Young, M. Gašić, S. Keizer et al., "The Hidden Information State model: a practical framework for POMDP-based spoken dialogue management," *Computer Speech and Language*, vol. 24, no. 2, pp. 150–174, 2010.

[9] S. Hara, N. Kitaoka, and K. Takeda, "Estimation method of user satisfaction using n-gram-based dialog history model for spoken dialog system," in *Proceedings of the LREC*, pp. 78–83, 2010.

[10] O. Lemon and l. Konstas, "User simulations for context-sensitive speech recognition in spoken dialogue systems," in *Proceedings of the EACL*, pp. 505–513, 2009.

[11] M. Rickert, M. E. Foster, M. Giuliani, G. Panin, T. By, and A. Knoll, "Integrating language, vision and action for human robot dialog systems," in *Universal Access in Human-Computer Interaction. Ambient Interaction*, pp. 987–995, 2007.

[12] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *Proceedings of the IEEE IRS/RSJ International Conference on Intelligent Robots and Systems (IROS '05)*, pp. 383–388, August 2005.

[13] T. Yonezawa, H. Yamazoe, A. Utsumi, and S. Abe, "Evaluating crossmodal awareness of daily-partner robot to user's behaviors with gaze and utterance detection," in *Proceedings of the 3rd ACM International Workshop on Context-Awareness for Self-Managing Systems (Casemans '09)*, pp. 1–8, May 2009.

[14] R. M. Maatman, J. Gratch, and S. Marsella, "Natural behavior of a listening agent," in *Proceedings of the Intelligent Virtual Agents*, vol. 3661 of *Lecture Notes in Computer Science*, pp. 25–36, 2005.

[15] S. Kopp, T. Stocksmeier, and D. Gibbon, "Incremental multimodal feedback for conversational agents," in *Proceedings of the Intelligent Virtual Agents*, vol. 4722 of *Lecture Notes in Computer Science*, pp. 139–146, 2007.

[16] L. P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 70–84, 2009.

[17] A. Kobayashi, K. Kayama, and E. Mizukami, "Evaluation of facial direction estimation from cameras for multi-modal spoken dialog system," in *Spoken Dialogue Systems for Ambient Environments*, pp. 73–84, 2010.

[18] O. Buß and D. Schlangen, "Modelling subutterance phenomena in spoken dialogue systems," in *Proceedings of Semdial*, pp. 33–41, 2010.

[19] S. E. Hudson, J. Fogarty, C. G. Atkeson et al., "Predicting human interruptibility with sensors: a Wizard of Oz feasibility study," in *Proceedings of the CHI New Horizons: Human Factors in Computing Systems*, pp. 257–264, April 2003.

[20] J. Begole, N. E. Matsakis, and J. C. Tang, "Lilsys: sensing unavailability," in *Proceedings of the Computer Supported Cooperative Work (CSCW '04)*, pp. 511–514, November 2004.

[21] S. Satake, T. Kanda, D. F. Glas, M. Imai, H. Ishiguro, and N. Hagita, "How to approach humans? Strategies for social robots to initiate interaction," in *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI '09)*, pp. 109–116, March 2009.

[22] N. Yankelovich, "How do users know what to say?" *Interactions*, vol. 3, no. 6, pp. 32–43, 1996.

[23] S. Benus, A. Gravano, and J. Hirschberg, "Pragmatic aspects of temporal accommodation in turn-taking," *Journal of Pragmatics*, vol. 43, no. 12, pp. 3001–3027, 2011.

[24] L. Ferrer, E. Shriberg, and A. Stolcke, "A prosody-based approach to end-of-utterance detection that does not require speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 608–611, April 2003.

[25] K. Laskowski, J. Edhund, and M. Heldner, "Incremental learning and forgetting in stochastic turn-taking models," in *Proceedings of the Interspeech*, pp. 2069–2072, 2011.

[26] K. Laskowski and E. Shriberg, "Corpusindependent history compression for stochastic turn-taking models," in *Proceedings of the ICASSP*, pp. 4937–4940, 2012.

[27] A. Raux and M. Eskenazi, "A finite-state turntaking model for spoken dialog systems," in *Proceedings of the Human Language Technologies*, 2009.

[28] R. Sato, R. Higashinaka, M. Tamoto, M. Nakano, and K. Aikawa, "Learning decision trees to determine turn-taking by spoken dialogue systems," in *Proceedings of the ICSLP*, pp. 861–864, 2002.

[29] J. J. Edlund and M. Nordstrand, "Turn-taking gestures and hourglasses in a multi-modal dialogue system," in *Proceedings of the IDS*, 2002.

[30] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. N. Nöth, "Desperately seeking emotions or: actors, wizards, and human

beings," in *Proceedings of the SpeechEmotion*, pp. 195–200, 2000.

[31] D. Litman and K. Forbes-Riley, "Spoken tutorial dialogue and the feeling of another's knowing," in *Proceedings of the SIGDIAL*, 2009.

[32] M. Goto, K. Itou, and S. Hayamizu, "A realtime filled pause detection system: toward spontaneous speech dialogue," in *Proceedings of the Eurospeech*, pp. 187–192, 1999.

[33] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. I511–I518, December 2001.

[34] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, 2003.

[35] E. J. E. J. Keogh and M. J. Pazzani, "A simple dimensionality reduction technique for fast similarity search in large time series databases," in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, pp. 122–133, 2000.

[36] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Tech. Rep., Department of Computer Science, National Taiwan University, 2003.

[37] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.