Estimating the User's State before Exchanging Utterances Using Intermediate Acoustic Features for Spoken Dialog Systems

Yuya Chiba, Takashi Nose, Masashi Ito, and Akinori Ito, *Member, IAENG*

***Abstract*—The spoken dialog system (SDS) is an example of a speech interface and has been included in several devices to help users operate the system. The SDS is beneficial for the user because it does not restrict the style of the user's input utterances, but sometimes makes it difficult to speak to the system. Conventional systems cannot give appropriate help to a user who does not make explicit input utterances since these systems have to recognize and parse a user's utterance in order to decide the next prompt. Therefore, the system should estimate the state of the user upon encountering a problem in order to start the dialog and provide appropriate help before the user abandons the dialog. Based on this assumption, we aim to construct a system which responds to a user who does not speak to the system. In this research, we defined two basic states of the user when the user does not speak to the system: the user is embarrassed by the prompt, or is thinking about how to answer the prompt. We discriminated these user states by using intermediate acoustic features and the facial orientation of the user. Our previous approach used several intermediate acoustic features determined manually, and it was not possible to discriminate the user's state automatically. Therefore, the present paper examines a method to extract intermediate acoustic features from low-level features, such as MFCC, $\log F0$, and zero cross counting (ZCC). We introduce a new annotation rule, and compare the discrimination performance with the previous feature set. Finally, the user's state was discriminated by using the combination of intermediate acoustic features and facial orientation.**

***Index Terms*—spoken dialog system, user's state, multi-modal information.**

I. Introduction

A spoken dialog system is a natural interface since speech commands are less subject to the physical constraints imposed by devices. On the other hand, a user must first know how to speak to control the system before using it, even though the system allows the user to speak anything. This requirement deters novice users or users who lack the motivation to converse with the system, and sometimes these users abandon the dialog with the system. To help a user who cannot start a dialog, the system must estimate the user's state before he/she gives up. Therefore, this paper focuses on the user's state after projecting the prompt and before the user makes an utterance. Here, we assumed the following three basic internal states when the user cannot make an input utterance:

Manuscript received August 6, 2015; revised September 26, 2015. This work was supported by JSPS Research Fellowships for Young Scientists Grand Number 263989.

Y. Chiba, T. Nose, and A. Ito are with the Graduate School of Engineering, Tohoku University, 6-6-05, Aramaki Aza Aoba, Aoba-ku, Sendai, 980-8579, Japan e-mail: yuya.spcom.ecei.tohoku.ac.jp

M. Ito is with the Tohoku Institute of Technology, 35-1 Yagiyama-Kasumicho, Taihaku-ku, Sendai-shi, Miyagi 982-8577, Japan

S_E: The state where the user does not know what to input and the user is embarrassed by the prompt.

S_C: The state where the user is considering how to answer the system's prompt.

S_N: The state where the user converses with the system smoothly.

Figure 1 shows an example of each state. The system can adapt to the user's problem by distinguishing the user's internal state. For example, the system can display a detailed message for the S_E user before he abandons the dialog, and can minimize the display of messages which irritate the S_C user.

Conventional systems take various approaches to help users who cannot start or maintain a dialog adequately, for example, "incremental prompt" which repeats the prompt at fixed intervals [1] or a mixed-initiative dialog system (i.e., [2], [3]). The incremental prompt is adopted by many dialog-based systems, but a help message could become troublesome for someone who uses the system regularly since the system responds to users uniformly. By contrast, mixed-initiative dialog systems can project the prompt adapted to the user by acquiring the user's information from the dialog history. However, these systems implicitly target users who can maintain a dialog with the system by themselves because the system has to exchange a few words. On the other hand, several researchers have studied how to address users who do not speak to the system. For instance, Satake et al. studied how to find an interaction target and how to approach the person in human-robot dialog [4], Hudson et al. examined a method to predict the willingness to be interrupted of a user in a working environment [5], and Michalowski et al. constructed a social robot that changes its behavior according to the degree of engagement of the interaction, which is estimated by the spatial location and attention of the user [6]. However, few studies have investigated the user's state as examined in this paper. In such case, if the user cannot make any input utterance, the system must identify the user's state from non-verbal information.

In our previous study [7], we also tried to estimate the state of the user by using the user's facial orientation and the length of acoustic events, such as filler, silence, and filled pause. In this approach, we obtained the harmonic mean between accuracy of S_E and S_c as 55.7%, but it was impossible to discriminate the user's state automatically since the length of the filler and silence were labeled manually. Therefore, this paper examines a method to estimate the intermediate features of the acoustic information by using low-level acoustic features, such as MFCC, $\log F0$, and zero cross counting (ZCC) for automatic discrimination of the

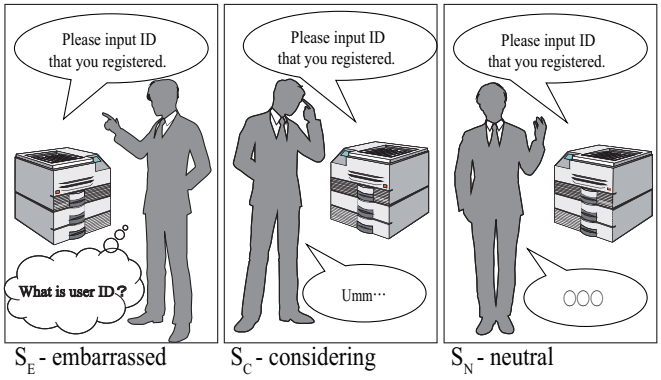

Fig. 1. User's state of stagnating dialog

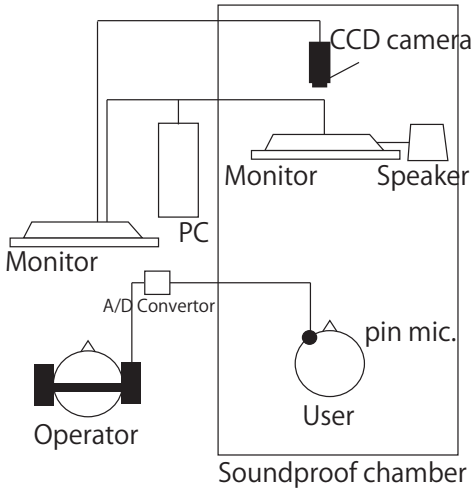

Fig. 2. Experimental set-up

user's state. Finally, we combine the acoustic features with the facial orientation and conduct a multi-modal discrimination to compare the performance with the result of the previous report.

II. Related works

Many researches have studied the user's internal state, such as the emotion [8], [9], [10], frustration [11], preference [12], paralinguistic information of the user's utterance [13], and familiarity with the task [14], [15]. In particular, "uncertainty" is related to our target internal state of the user. Forbes-Reley and Litman [8] and Pon-Barry et al. [16] investigated the user's uncertainty, and Forbes-Reley and Litman introduced a framework for estimating the user's uncertainty in a tutor system, and reported that projecting the prompt corresponding to the user's uncertainty enhances the learning effect [17]. These studies focused on the user's state after beginning the dialog (i.e. after the user's response to the system's first utterance), and indicated that the linguistic information of the utterance is efficient for estimating the user's state.

Another phenomenon of the dialog actively studied is turn-taking [18]. How to hold and yield the floor is important for the speech interface, such as IVR, to converse with the user naturally [19], [20]. Morency et al. studied a method to predict the listener's back-channels from human-human conversation [21], Chen and Harper investigated the shifting of floor control [22], and Kopp et al. proposed a model for generating feedback [23]. The linguistic, prosodic, and visual information determine the timing of the turn switches, backchannels and interruptions. Jurafsky et al. examined the Switchboard Corpus, and indicated that lexical knowledge plays a role in distinguishing backchannel acts, such as continuers, assessments, or incipient speakership [24]. They also showed that the identification of some backchannel acts is affected by the prosodic cues. Koiso et al. showed the syntactic and prosodic features are related to turn-taking and backchannels by analyzing the dialog between Japanese [25]. Compared with these related works, we cannot exploit linguistic cues of the utterance because we need to determine the user's state before his/her utterance. Another difference between our work and the other works is that the goal of study is not to decide the timing of presenting the system's response but to estimate the user's state.

The estimations of the states S_E and S_C were related to "Feeling of Another's Knowing (FOAK)" [26] in the dialog between humans. Brennan and Williams investigated how people reliably determine other people's "Feeling of Knowing (FOK)," which is the partner's feeling that he would recognize the correct answer to the question. They concluded that the rate of FOAK is affected by the interlocutor's latency to respond to the question, the presence of filler, and surface expression of the utterance from the analysis of question-answering dialog between humans. In addition, Swerts and Krahmer studied multi-modal dialog data, and listed gaze acts, eyebrow movements, and facial expression as visual cues [27]. The state S_E corresponds to the state where the user's FOK is low, and other states correspond to the state where the user's FOK is relatively high. Therefore, this non-verbal information seems to be efficient for estimating the target user's state.

III. Dialog data collection and rating

A. Recording dialog between the system and the user

Firstly, we collected the dialog data because the amount of data was small in previous study [7]. The experimental data were collected on the Wizard-of-Oz basis [28]. In the WOZ method, as much natural dialog data as possible is collected by having the user converse with a simulated dialog system. Figure 2 shows the experimental circumstance. The experiments were conducted in a soundproof chamber. To record the speech and video of the user's frontal face, a CCD camera was installed above a display screen in front of the user, and the user wore a lapel microphone. Additionally, an agent with a simple cartoon-like face was projected on the display to keep the user's attention. The agent was controlled by an experimenter outside of the chamber according to the user's utterance. The audio signals were recorded in PCM format at 16 kHz sampling and 16-bit quantization. The recorded video clips were stored as AVI files with 24-bit color depth, 30 frames/s. We implemented a question-and-answer task in which the system posed questions and the user answered them. The questions were about common knowledge or numbers memorized in advance, such as, "Please input today's date." and "Please input your ID." Sixteen users (14 males and 2 females) participated in the dialog collection. If no speech occurred because the user was thinking of what to input or was embarrassed by the prompt,

TABLE I
EVALUATION RESULTS OF USER'S STATE

S_E	S_C	S_N	Total
59 (7.5%)	195 (24.6%)	538 (67.9%)	792

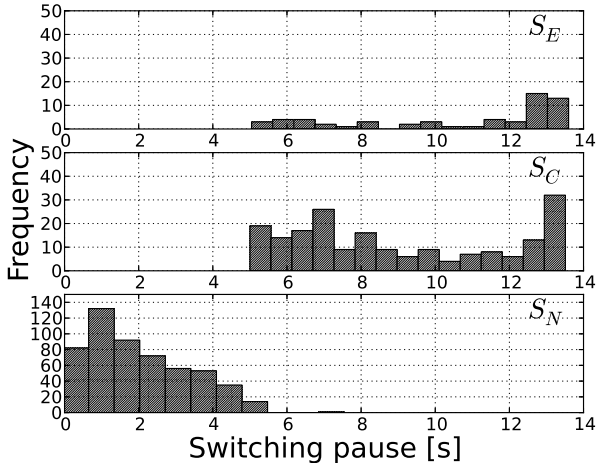

Fig. 3. Distribution of switching pause

the system repeated the same question every 15 seconds. In addition, the users were permitted to utter "I don't know," but to the minimum extent possible.

The recorded clips were divided into sessions, where one session included one interchange of the system's prompt and the user's response.

B. Annotation of user's state

After collecting the dialog data, we prepared video clips for labeling the sessions, which contained only the non-verbal behavior of the user. To do this, we first masked the system's prompt by a tone signal (the evaluator could observe only the user's face), and truncated the last part of the clips that contained the user's input utterance. In previous experiments [7], sessions which had a short switching pause tended to be classified as S_N; therefore, we used only the session whose switching pause was longer than 5 seconds for annotation, and the other sessions were annotated as S_N. We employed five evaluators to annotate each video clip as state S_E, S_C or S_N, and took the majority vote of the evaluators' decisions to determine the state of the user of a session. The evaluation results are shown in Table I. Fleiss' κ among the evaluators was 0.22 (fair agreement), where sessions with a tie vote were excluded.

C. Distribution of switching pause

Figure 3 shows the distribution of the switching pause of S_E, S_C, and S_N, where the switching pause is defined as the length of the segment between the end of the system's prompt and the beginning of the user's input utterance. As Figure 3 shows, the sessions of S_N and the others are clearly separated by the length of the switching pause. In fact, a Naive Bayes classifier using switching pause as the feature discriminated the session of S_N and the others with an accuracy of over 99.0%. On the other hand, it is difficult to classify the sessions of S_E and S_C because they have almost the same distribution of switching pause.

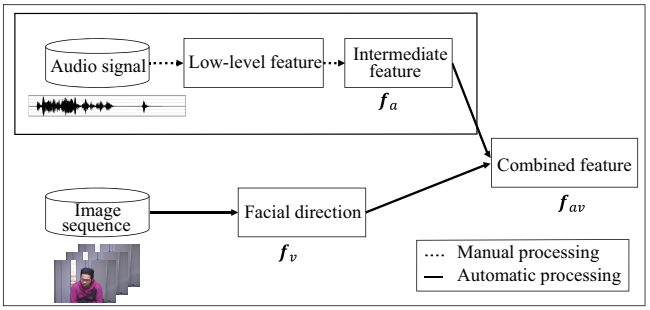

Fig. 4. Feature extraction of the previous experiment

IV. DISCRIMINATION OF USER'S STATE USING INTERMEDIATE ACOUSTIC FEATURES

Next, we conduct discrimination using features with fixed dimensions to select the most appropriate acoustic feature set. In the previous paper, we discriminated the user's state by using audio-visual features. Although the visual features and part of the acoustic features were obtained automatically, several acoustic features had to be determined manually; therefore, automatic discrimination of the user's state was not possible. Figure 4 shows the flow of the feature determination of the previous paper. To discriminate the user's state, we used the symbols of acoustic events labeled frame-by-frame, which we called "intermediate acoustic features." Figure 5 shows an example of the sequence of the user's states and intermediate acoustic features. The upper row of the figure denotes the user's state and the middle row shows the intermediate acoustic features.

The feature set of the previous report [7] is denoted as S1 (manual). In this study, we prepared a novel inventory of the intermediate features to simplify the estimation because the previous selection was controversial. The manually annotated intermediate features belonging to the novel inventory were denoted as S2 (manual) and intermediate features estimated by the neural network were denoted as S2 (NN).

A. Kinds of intermediate acoustic features

First, we summarized the set of intermediate features in Table II. S1 includes the length of the following four acoustic events.

Length of user's filler (AF):
: This feature represents the duration of the filler segment of the user.

Length of silence segment (ASI):
: This feature represents the length of the unvoiced segment.

Length of filled pause (AFP):
: This feature represents the duration of the filled pause.

Length of switching pause (ASP):
: This feature represents the length of switching pause between the system's utterance and the user's input utterance.

From the unpaired t-test, a significant difference ($p \leq 0.05$) was observed at AF and ASI between S_E and S_C, and thus we selected them as S1. In addition, we employed a filled pause as a feature in S1. In the filled pause segment, the vowels tend to be lengthened, and such pauses are frequently

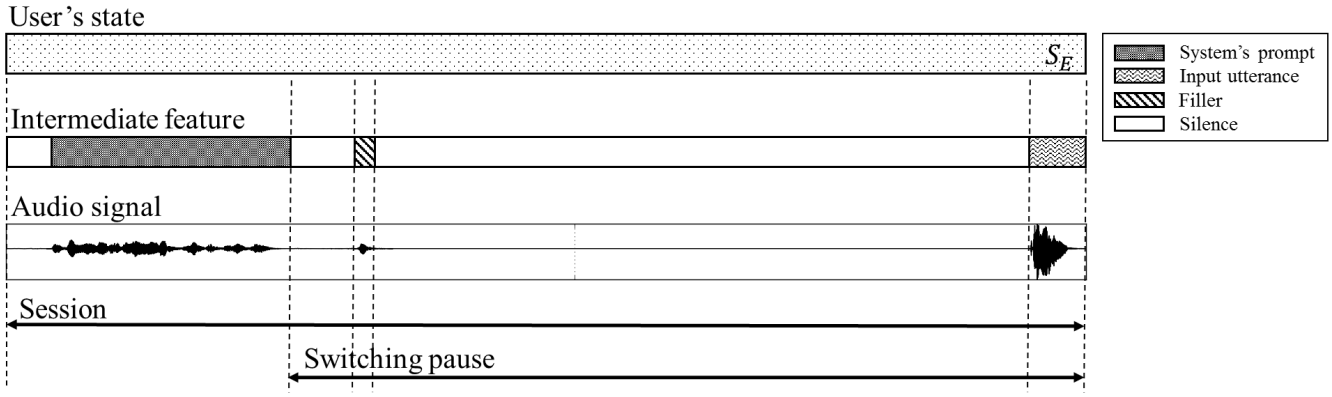

Fig. 5. Example of the user's state and intermediate acoustic features

TABLE II
INVENTORY OF INTERMEDIATE FEATURES

Symbol	Description	Feature set	
		Conventional (S1)	Proposed (S2)
AS	system's utterance		✓
AI	user's utterance		✓
AF	user's filler	✓	✓
AB	user's breath		✓
ASE	user's self speech		✓
AW	user's whisper speech		✓
ASI	unvoiced segment	✓	✓
AFP	user's filled pause	✓	$*$
ASP	switching pause	✓	$*$
AR	user's repair utterance		$*$
AO	other speech events of S1 (i.e., ASE + AW)		$*$

$*$ denotes that the set includes the feature indirectly. S2 contains information of switching pauses because the sum of the intermediate features excluding AS and AI equals ASP. Similarly, S2 includes AR as AI, and AO as ASE or AW.

observed in utterances when the user is thinking (which corresponds to S_C). Goto et al. proposed a method to detect filled pauses and described that the filled pause serves to maintain the speaker's turn and to express the mental state while thinking of the next utterance [29]. In contrast, we used ASP for discriminating S_N from the other states (i.e. S_E and S_C).

Because the set S1 involves both manually-annotated and automatically-annotated features, we prepared a new annotation criterion, S2, to simplify the estimation of the intermediate features. First of all, the user's filled pause was integrated with AF because filled pauses and fillers have similar physical characteristics. In addition, the user's repair speech (AR) was combined with the input utterance (AI) because this event tends to occur just before the input utterance. On the other hand, other speech events (AO) were separated into whisper speech (AW) and self speech (ASE) for analysis because the occurrence of the user's whisper speech or self speech seemed to affect the evaluator's decision from the evaluator's self reports. The details of each feature are described below.

Length of system's utterance (AS):
: This feature represents the duration of the system's utterance.

Length of user's input utterance (AI):
: This feature represents the duration of the user's input utterance.

Length of user's filler (AF):
: This feature represents the duration of the filler segment of the user.

Length of user's breath (AB):
: This feature represents the duration of the breath segment of the user.

Length of user's self speech (ASE):
: This feature represents the duration of the user's speeches other than input utterances.

Length of user's whisper speech (AW):
: This feature represents the duration of the user's whisper speech.

Length of silence segment (ASI):
: This feature represents the length of the unvoiced segment.

Because the sum of the intermediate features other than AS and AI equals ASP, S2 indirectly contains information of switching pauses.

B. Estimation of intermediate acoustic features

Neural networks were used for estimating the intermediate features. The low-level acoustic features were extracted from the audio signal and used as the input of the neural network. The outputs of the neural network are scores of the acoustic events. Estimation of the intermediate acoustic features was conducted frame-by-frame, then the estimation results were accumulated to represent the length of each event. This section summarizes the estimation process.

1) Low-level acoustic features: To represent the spectral characteristics of the speech, MFCC was employed as the low-level acoustic feature. We used a 39-dimension MFCC including the velocity and acceleration coefficient of the

TABLE III
CONDITIONS OF LOW-LEVEL ACOUSTIC FEATURE EXTRACTION

	MFCC	log $F0$	ZCC
Window length	25.0 ms	17.0 ms	10.0 ms
Frame shift	10.0 ms	10.0 ms	10.0 ms

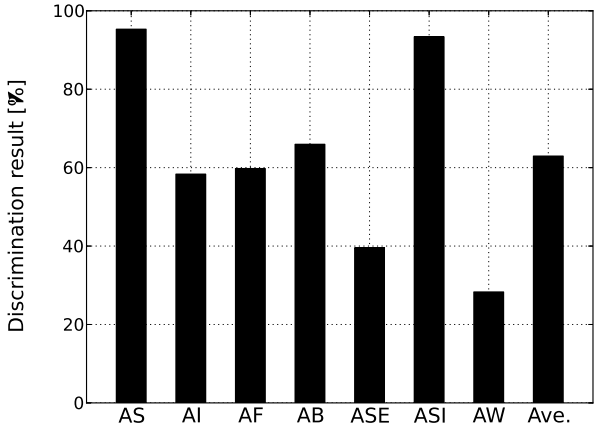

Fig. 6. Estimation results of acoustic events

lower 12th-order coefficients and log power. In addition, a differential component of log $F0$ was used to represent the prosodic feature of the speech, and zero cross count (ZCC) was used to distinguish voiced and unvoiced segments. Therefore, the total number of dimensions of the audio features was 41. The basic conditions for extracting each feature are shown in Table III. Here, five frames (the current frame, the two previous frames and the two following frames) were used to calculate the Δ and $\Delta\Delta$ components of MFCC, and the Δ component of log $F0$.

2) Training of the classifier: In the training stage, three-layered neural networks having input, hidden and output layers including the bias unit were trained. We employed a softmax activation function at the output layer in order to obtain the outputs as the probability of the acoustic events. The activation function of the hidden layer was a logistic sigmoid function. The number of units of hidden layers n_h and the number of training epochs n_e were comprehensively changed to identify the optimum parameters. n_h was changed from 10 to 100 and n_e was changed from 1 to 100. The experiment was conducted based on 5-fold cross-validation.

Table IV shows the total number of frames of each intermediate acoustic feature. As shown in the table, the frequency of acoustic events was not balanced, and ASI, AS, and AI accounted for the majority.

3) Estimation results: The frame-by-frame estimation result was chosen from the equation:

$$\hat{c}_t = \arg\max_c p_{tc} \tag{1}$$

where p_{tc} is the score of the intermediate acoustic feature c and $\hat{c}_t$ is the intermediate feature estimated at time t. The result of the estimation is shown in Figure 6. The most accurate estimation result was obtained when $n_h = 90$ and $n_e = 3$, and then the average estimation ratio was 62.95%.

As the results show, AS and ASI obtained a high estimation accuracy. Since these intermediate features have a distinct characteristic, the classification was considered to be easy. On the contrary, ASE and AW are difficult to discriminate: ASE tended to be classified as AI or AF, and AW tended to be taken as several other acoustic events. One of the reasons for the low classification ratio is that the low-level features did not capture the characteristics of these acoustic events due to the shortage of training data. It is considered that ASE has the same characteristics as some voiced classes (AI or AF), and that AW has the same characteristics as breathy voice. Also, there were insufficient data for AF and AB, but these classes maintained a reasonable discrimination ratio (around 60.0 $\%$), which is why the characteristics of these features have little variation.

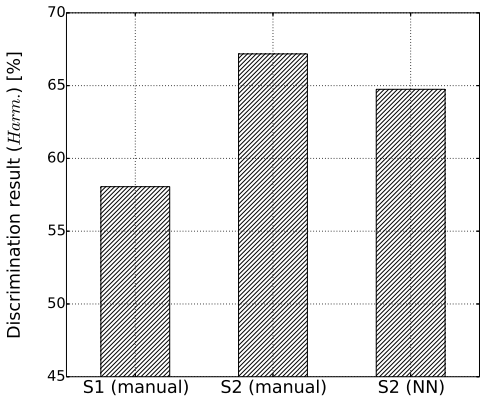

Fig. 7. Discrimination results using audio features

In addition, the method that evaluates the audio signal frame-by-frame cannot adequately capture the temporal variation of the audio signal because the low-level features tend to be unstable at points where the intermediate acoustic features change, so the performance of the estimation declined at the start and end of the events. Therefore, it is necessary to examine other estimation methods to improve the performance (e.g. GMM or HMM).

4) Length of the intermediate features: Finally, we constructed a feature vector for discriminating the user's states by using the estimation results. This was achieved by accumulating the estimated intermediate acoustic features of each frame session by session. The accumulated intermediate acoustic features correspond to the duration of each acoustic event. The intermediate acoustic feature estimated by the neural network was represented as a 7-dimensional vector and denoted as S2 (NN).

C. Experimental conditions for discriminating the user's state

The user's state was discriminated by the Support Vector Machine (SVM) with RBF kernel. The hyperparameters of the classifier were decided by grid-searching. As mentioned in the previous section, the session of S_N and the other states were clearly distinguished by the duration of the session, therefore we used only the sessions of S_E and S_C for the following experiments. Hence, each experiment was a two-class discrimination task. Here, the total accuracy tends to increase as the determined class tends toward S_C because the amount of data is not uniformly distributed (see Table I); therefore, the harmonic mean (denoted as H) was employed

TABLE IV
FREQUENCY OF ACOUSTIC EVENTS

AS	AI	AF	AB	ASE	ASI	AW	Total
169645	70958	9887	4427	8612	317756	5664	586949
(28.90%)	(12.09%)	(1.68%)	(0.75%)	(1.47%)	(54.14%)	(0.96%)	

for measuring the performance, and was calculated by the equation:

$$H = \frac{2C_E C_C}{C_E + C_C} \quad (\%) \tag{2}$$

where, C_E and C_C represent the discrimination correctness of states S_E and S_C calculated by equations (3) and (4):

$$C_E = \frac{N'_E}{N_E} \times 100.0 \quad (\%) \tag{3}$$

$$C_C = \frac{N'_C}{N_C} \times 100.0 \quad (\%) \tag{4}$$

N_E and N_C respectively denote the number of sessions of S_E and S_C, and N'_E and N'_C are the number of sessions discriminated correctly by the classifier. The experiments were conducted by 5-fold cross-validation using same split to the intermediate feature estimation. In addition, all of the feature vectors used in the following section were L2-normalized.

D. Results of discriminating the user's state by using acoustic feature

Figure 7 shows the discrimination results when using the acoustic features mentioned above. As shown, the result of S2 (manual) was higher than that of S1 (manual). Therefore, the feature set prepared in this paper is more efficient than the previous feature set for discriminating the user's state, and suggests the appropriateness of the novel inventory of the feature set. On the other hand, S2 (NN), which is obtained by the estimation, does not surpass the ideal feature set S2 (manual). The decrease in performance seems to be due to the noise produced by the estimation error of the intermediate features, indicating that the accuracy of estimating the intermediate features is important to improve the definitive results of discriminating of the user's state.

V. USER'S STATE DISCRIMINATION USING AUDIO-VISUAL FEATURES

In our previous work [7], we used visual features in addition to acoustic features. Therefore, we tried to discriminate the user's state by using audio-visual features. We employed the facial orientation as the visual feature, the same as in the previous work. The facial orientation was calculated from the relative positions of the eyes, nose and face of the user, and was represented by 3-directional angles (i.e. yaw, roll and pitch).

A. Detection of main facial components

The facial components used to calculate head movement were detected as follows. First, the feature points of the face were extracted by the Constraint Local Model (CLM) [30]. In this method, a model of the feature points is fitted after detecting the facial region from the whole image in the frame. Figure 8 shows a model of the feature points and

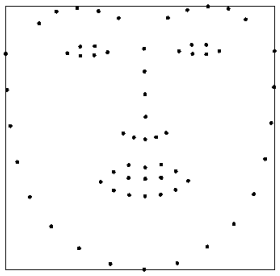

Fig. 8. Model of facial feature points

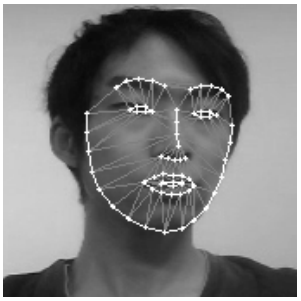

Fig. 9. Result of feature extraction

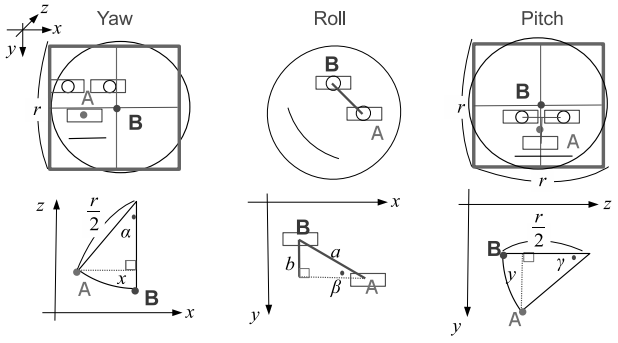

Fig. 10. Calculation of face orientation feature

Figure 9 is an example of the result of fitting. Then, the main facial components of the user such as the left eye, right eye and nose were detected based on the results of feature point extraction. Although more complicated facial movements of the user can be captured from the facial feature points, we used only the facial angles, the same as in our previous work, to verify the performance of the acoustic feature set prepared in this study. We will examine the selection of facial features in a future study.

B. Calculation of facial orientation

Three-dimensional facial orientation was calculated based on the locations of the eyes, nose and face. We approximated the shape of the human head as a sphere and used a sine value instead of the angle (Figure 10).

$$yaw \equiv \frac{2x}{r} = \frac{2(A_x - B_x)}{r} \tag{5}$$

$$roll \equiv \frac{b}{a} = \frac{B_y - A_y}{\sqrt{(B_x - A_x)^2 + (B_y - A_y)^2}} \tag{6}$$

$$pitch \equiv \frac{2y}{r} = \frac{2(A_y - B_y)}{r} \tag{7}$$

where, r is the diameter of the head, estimated as the width of the face region. The points $\boldsymbol{A} = (A_x, A_y)$ and $\boldsymbol{B} = (B_x, B_y)$ are feature points in a frame, as shown in Table V.

TABLE V
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

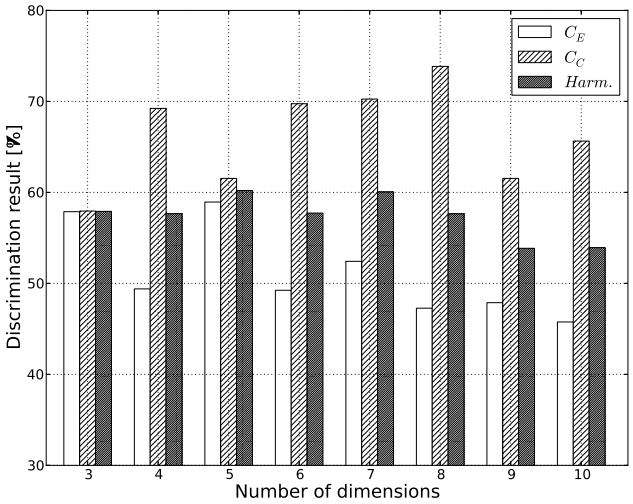

Fig. 11. Discrimination results using facial orientation

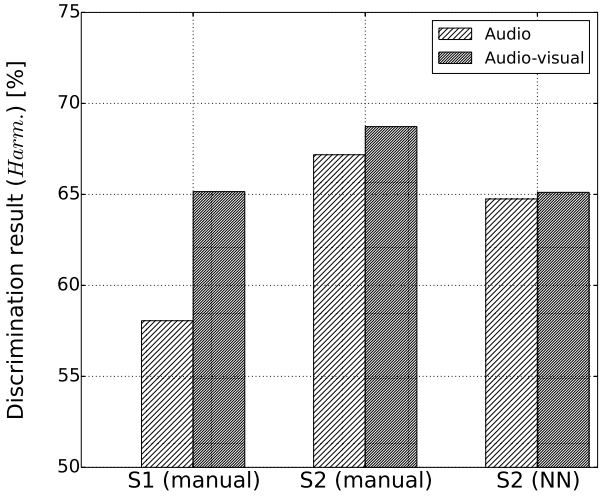

Fig. 12. Discrimination results using audio-visual features

We calculate yaw, roll and pitch frame-by-frame, and denote these values at frame t as $yaw(t)$, $roll(t)$ and $pitch(t)$.

The length of facial orientation of each session depends on the duration of the session, but the feature vectors should have a fixed dimension to simplify the calculation of discrimination function. The facial orientations were compressed linearly to normalize the length. In this paper, the compressed feature vectors $\bar{x}_1, \cdots, \bar{x}_n$ are calculated from the facial direction sequence $x_1, \cdots, x_N$ $(n \leq N)$ by the equation:

$$\bar{x_i} = \frac{n}{N} \sum_{j=(N/n)(i-1)+1}^{(N/n)i} x_j \tag{8}$$

$yaw(t)$, $roll(t)$ and $pitch(t)$ were compressed to 50 samples.

C. Dimension reduction of the facial orientation

Each session has 150 (50×3) dimensions of the facial orientation feature by linear compression. The visual features still have large dimensions in contrast with the acoustic features. Accordingly, the dimensionality of facial orientations was reduced by principal component analysis (PCA). The PCA was conducted based on the 5-fold cross-validation using same split to the estimation of the intermediate acoustic feature and discrimination of the user's state. We reduced the dimension of the facial orientation of each fold by projecting the principal axes obtained from other four folds.

Figure 11 shows the result of discriminating the user's state by using only compressed visual features. The horizontal axis represents the number of dimensions after the reduction. The correctness of the discrimination was always around 60.0%, so we used the 3 dimensional facial orientation for audio-visual discrimination.

D. Combining the audio and visual feature

The audio and visual features were combined simply as:

$$f_{av} = (f_a, f_v) \tag{9}$$

where, f_a is the intermediate acoustic feature vector, f_v is the visual feature vector, and f_{av} is the combined vector.

E. Discrimination results using audio-visual features

Discrimination experiments were conducted based on the 5-fold cross-validation based on the same split to the above-mentioned experiments. Figure 12 shows the results of using audio-visual feature f_{av}. The best result for each condition obtained by grid-searching is shown in the figure. The figure also shows the result using only the acoustic features for comparison. These are the same as the value indicated in Figure 7. As shown, the discrimination results were improved by appending the visual features. Especially, the feature set S1 (manual) shows a large improvement by including the visual feature. This results shows that the S1 (manual) and facial orientation feature were efficiently combined as indicated in our previous paper [7]. However, improvement of the discrimination performance of S2 (manual) and S2 (NN) was not large (around 1.0 or 2.0 points), indicating that we need to select more appropriate visual features to combine with the new feature set to improve the discrimination of the user's state. The evaluators of the dialog reported that the variation of the user's facial expression and gaze direction had affected their judgment; therefore, it is necessary to select novel visual features representing these characteristics. We can extract these features by conducting several other computations in addition to the extraction of facial feature points; we will examine visual features in more detail in a future study.

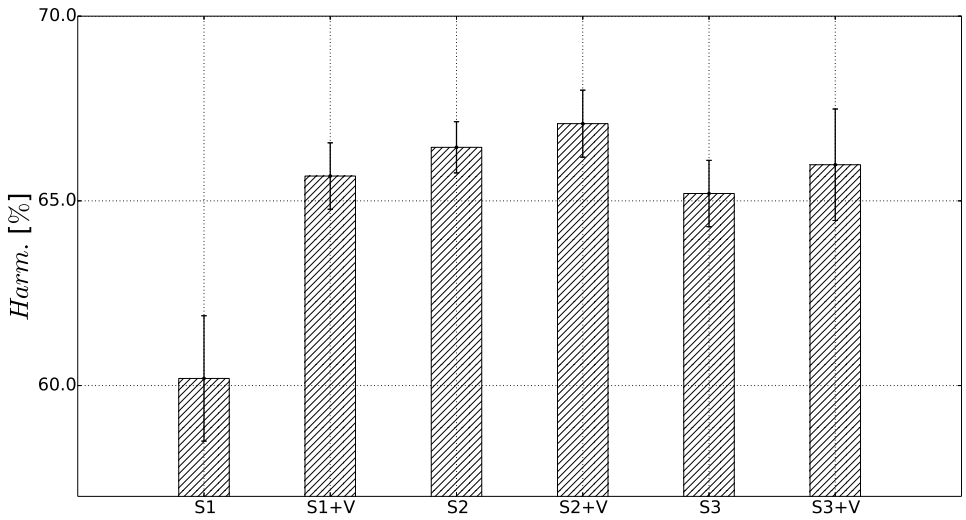

Fig. 13. Discrimination result obtained from repeated experiments

In addition, we investigated the performance of the features by repeating the discrimination experiments. The experiment was 5-fold cross-validation test, which uses the folds randomly assigned from the data set, and repeated 10 times. Figure 13 shows the discrimination results. The figure shows the average harmonic means of the discrimination ratio and the error-bars indicate the standard deviation of them. We conducted one-way repeated measures ANOVA factoring the feature set to validate the statistical significance of the difference between the total discrimination rates. As a result, the significant difference was observed among the method ($F = 41.4$, $p \leq 0.001$). From the Bonferroni multiple comparison, we obtained the 1% significant difference between S1 and the other feature set. Thus, the efficacy of the novel acoustic feature set to the discrimination was indicated stochastically. These results also indicate that combining the visual information is especially efficient to S1, and the efficacy to the other feature set is limited.

VI. Conclusion

In this research, we tried to discriminate the user's state in a spoken dialog system when the user does not make input utterance. We tried to extract the intermediate acoustic features automatically because we determined several acoustic features manually in our previous study. A novel set of intermediate features was estimated by a neural network using MFCC, F0 and ZCC of the speech. The result of estimating the intermediate acoustic features was 62.95% on average. The definitive user's state was discriminated by SVM. When discriminating the user's state using only acoustic features, the correctness was high in order of S2 (manual), S2 (NN), and S1 (manual). Therefore, the novel feature set S2 (manual) was more efficient than the previous one (S1). On the other hand, the result deteriorated when using the estimated intermediate feature (S2 (NN)). This result indicates that the performance of estimating intermediate features involves a definitive result, so we need to examine the estimation method carefully. Then, we combined the audio-visual features and discriminated the user's state. We used facial orientation as a visual feature, the same as in our previous research, which improved the correctness of discrimination. Especially, the improvement of the result of S1 (manual) was large because this audio-visual feature was selected appropriately in the previous paper. However, there was little improvement in the correctness of S2 (manual) and S2 (NN). Therefore, we need to search for another visual feature which can be combined with them efficiently. In a future work, we will use the user's eye movement or facial expression extracted from the facial feature points for the discrimination. Another problem is that the entire discrimination ratio was not adequate (the harmonic mean between the accuracy of S_E and S_c was below 70.0%). We also need to examine a discrimination algorithm for enhancing the performance.

References

[1] N. Yankelovich, "How Do Users Know What to Say?," *Interactions*, vol. 3, no. 6, pp. 32–43, 1996.

[2] G. Chung, "Developing a Flexible Spoken Dialog System Using Simulation," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 63–70, 2004.

[3] D. Bohus and A. I. Rudnicky, "The RavenClaw Dialog Management Framework: Architecture and Systems," *Computer Speech & Language*, vol. 23, no. 3, pp. 332–361, 2009.

[4] 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*, pp. 109–116, 2009.

[5] S. Hudson, J. Fogarty, C. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. Lee, and J. Yang, "Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 257–264, 2003.

[6] M. P. Michalowski, S. Sabanovic, and R. Simmons, "A Spatial Model of Engagement for a Social Robot," in *Proceedings of 9th IEEE International Workshop on Advanced Motion Control*, pp. 762–767, 2006.

[7] Y. Chiba and A. Ito, "Estimating a User's Internal State before the First Input Utterance," *Advances in Human-Computer Interaction*, vol. 2012, p. 11, 2012.

[8] 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, no. 9, pp. 1115–1136, 2011.

[9] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive Learning for Enhanced Audiovisual Emotion Classification," *IEEE Trans. on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[10] C. M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of Negative Emotions from the Speech Signal," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 240–243, 2001.
[11] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based Automatic Detection of Annoyance and Frustration in Human-computer Dialog.," in *Proceedings of INTERSPEECH*, pp. 2037–2040, 2002.
[12] 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, pp. 329–351, 2004.
[13] S. Fujie, Y. Ejiri, K. Nakajima, Y. Matsusaka, and T. Kobayashi, "A Conversation Robot Using Head Gesture Recognition as Para-linguistic Information," in *Proceedings of 13th IEEE International Workshop on Robot and Human Interactive Communication*, pp. 159–164, 2004.
[14] K. Jokinen and K. Kanto, "User Expertise Modelling and Adaptivity in a Speech-based E-mail System," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 88–95, 2004.
[15] 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.
[16] H. Pon-Barry, K. Schultz, E. O. Bratt, B. Clark, and S. Peters, "Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems," *International Journal of Artificial Intelligence in Education*, vol. 16, no. 2, pp. 171–194, 2006.
[17] K. Forbes-Riley and D. Litman, "Designing and Evaluating a Wizarded Uncertainty-adaptive Spoken Dialogue Tutoring System," *Computer Speech & Language*, vol. 25, no. 1, pp. 105–126, 2011.
[18] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-taking for Conversation," *Language*, vol. 50, pp. 696–735, 1974.
[19] V. H. Yngve, "On Getting a Word in Edgewise," in *Proceedings of 6th Meeting of Chicago Linguistics Society*, pp. 567–578, 1970.
[20] A. Gravano and J. Hirschberg, "Turn-taking Cues in Task-oriented Dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.
[21] L.-P. Morency, I. de Kok, and J. Gratch, "Predicting Listener Backchannels: A Probabilistic Multimodal Approach," in *Intelligent Virtual Agents*, pp. 176–190, 2008.
[22] L. Chen and M. P. Harper, "Multimodal Floor Control Shift Detection," in *Proceedings of the 2009 International Conference on Multimodal Interfaces*, pp. 15–22, 2009.
[23] S. Kopp, T. Stocksmeier, and D. Gibbon, "Incremental Multimodal Feedback for Conversational Agents," in *Intelligent Virtual Agents*, pp. 139–146, 2007.
[24] D. Jurafsky, E. Shriberg, B. Fox, and T. Curl, "Lexical, Prosodic, and Syntactic Cues for Dialog Acts," in *Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*, pp. 114–120, 1998.
[25] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An Analysis of Turn-taking and Backchannels based on Prosodic and Syntactic Features in Japanese Map Task Dialogs," *Language and Speech*, vol. 41, no. 3-4, pp. 295–321, 1998.
[26] 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," *Journal of Memory and Language*, vol. 34, no. 3, pp. 383–398, 1995.
[27] M. Swerts and E. Krahmer, "Audiovisual Prosody and Feeling of Knowing," *Journal of Memory and Language*, vol. 53, no. 1, pp. 81–94, 2005.
[28] Y. Chiba, M. Ito, and A. Ito, "Effect of Linguistic Contents on Human Estimation of Internal State of Dialog System Users," in *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*, pp. 11–14, 2012.
[29] M. Goto, K. Itou, and S. Hayamizu, "A Real-time Filled Pause Detection System for Spontaneous Speech Recognition," in *Proceedings of Eurospeech*, pp. 227–230, 1999.
[30] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable Model Fitting by Regularized Landmark Mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.