Effect of Linguistic Contents on Human Estimation of Internal State of Dialog System Users

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

[1]Graduate School of Engineering, Tohoku University, Sendai, Japan
[2]Department of Electronics and Intelligent System, Tohoku Institute of Technology, Sendai, Japan
yuya@spcom.ecei.tohoku.ac.jp, itojin@tohtech.ac.jp, aito@spcom.ecei.tohoku.ac.jp

Abstract

We have studied estimation of dialog system users' internal state before the input utterance. In a practical use of a dialogue-based system, a user is often perplexed with the prompt. An ordinary system provides more detailed information to the user taking time to input, but the help is meddlesome for the user considering the answer to the prompt. To make an appropriate response, the spoken dialogue system needs to consider the user's internal state before the user's input. In the previous paper, we proposed a method for estimating the internal state using multi-modal cues; however, we did not separate effects of several factors (e.g. linguistic information of the prompt, visual information, and acoustic information). Thus, it was not clear which factor affects the evaluation of the dialog session to what extent. In this paper, we examined more detailed evaluation by human evaluators, separating linguistic contents (the system's prompt utterance and the user's reply utterance) from the other non-verbal behavior, and assessed the effect of the linguistic contents on the estimation of the user's internal state.

Index Terms: multi-modal interface, user modeling, non-verbal information

1. Introduction

A spoken dialog system is desired to respond to a user flexibly. User modeling and estimation of the user's state [1] is an important issue for realizing flexible spoken dialog systems. There have been many works on this issue so far, where most researches focus on estimation of the internal states in the dialog [2, 3] or before the dialog [4].

These researches assume that the user makes an answer immediately when the dialog system gives a prompt message. However, users under actual environment do not make the input on occasion. For instance, the user could abandon the session without uttering a word if the user did not understand the meaning of the system's prompt; or the user could take a time to answer when considering how to answer the prompt.

In our previous work, we focused on the user's behavior after listening to the system's prompt and before answering the prompt [5]. User modeling at this phase (before the user's first utterance) is important because we can recognize users who have difficulty understanding the system's prompt and making answer to the system. In this work, we exploited audio-visual features of the user's behavior such as the duration from the prompt to the user's answer, length of the users filled pause and silence, and the user's face orientation.

All the features examined in the previous work were extracted only from the observation of the user's behavior before the answer utterance. However, when labeling the dialog data, human annotators watched whole of the dialog session from the system's prompt to the end of the user's answer, which included additional linguistic information that was not used for automatic estimation, such as the system's prompt and the user's input utterance. Therefore, it is not clear how linguistic information affected the annotator's judgment.

In this paper, we conduct more detailed evaluation experiments. In this experiment, we created video clips of four different conditions that contained different information, and asked the evaluators to judge the user's internal state by watching the video. Then we compared judgments of different condition to assess the importance of linguistic information included in the system prompt and the user's input utterance, as well as audio information before the user's input utterance.

2. Internal states of a user before the first utterance

In a human-human dialogue, interlocutors converse more or less estimating the dialog partner's internal state. Here, we defined three internal states of a user about to answer to a dialog system [5]. 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. Estimation of these internal states will help the system to generate additional prompt when the user does not reply to the system within a certain duration.

Human estimation of dialog partner's internal state is based on feeling that another person to be knowing an answer to the question (in other words, whether other interlocutors could respond to his/her utterance or not). This is called "Feeling of Another's Knowing (FOAK)" [6]. A correlation between audio-visual cues and FOAK was also investigated [7]. This work used linguistic information of the user's input utterance, which means that this method cannot be used for the above purpose because the estimation needs the user's utterance.

3. Linguistic contents and estimation of the internal state

As stated, relationship between linguistic contents and estimation of the internal state is unclear. Therefore, we addressed the following three questions:

(Q1) How does the answer from the user affect the decision of the evaluators?
(Q2) How does the content of the question (the system's prompt) affect the decision of the evaluators?
(Q3) Is audio information really useful for the decision?

We used video clips of users who were making conversation with a dialog system. One clip contained one session, where the system gave a prompt utterance, and the user answered. We split the clip into audio and visual parts, and temporally divided into three parts, as follows.

1. Audio (A1) and video (V1) of the system's prompt utterance gave to the user
2. Audio (A2) and video (V2) after the prompt and before the user's answer
3. Audio (A3) and video (V3) of the user's answering utterance

To investigate the above three questions, we created the following four kinds of clips, and carried out experiments to compare judgments of the subjects with different kinds of clips.

Clips A: Clips with all information (V1, V2, V3, A1, A2, A3)

Clips B: Clips without the answer from the user (V1, V2, A1, A2)

Clips C: The system's prompt utterance of Clips B was substituted with tone signal (V1, V2, A2)

Clips D: The audio signal of Clips C was removed (V1, V2)

We can investigate the answer of (Q1) by comparing judgments for Clips A and B. Then we investigate (Q2) by comparing the result of Clips A and B with that of Clips C. Finally, we compare with these results with the result of Clips D for answering (Q3).

4. Collection of dialog data

We collected the dialog data on the Wizard-of-Oz (WOZ) basis, where subjects make dialogs with a dialog system controlled by a human operator. We prepared a "Question-and-answer" task for the dialog, where the system asks the subject a question and the subject answers it. We prepared 44 patterns of system questions.

We prepared an agent on the LCD monitor to keep the subjects' attention. The agent is a simple cartoon-like face, which were controlled by the operator. The operator gave the system's prompt and reply to the subject using a speech synthesizer.

The experiment was conducted in a sound-proof chamber. The system utterance was played by a speaker connected to the PC. The operator stayed outside of the chamber and controlled the agent remotely. The subjects wore a lapel microphone. A CCD camera was installed above the monitor to record the frontal face of the subject during a dialog. The operator could monitor both the speech and video of the subject from outside the chamber.

We employed 16 subjects (14 males and 2 females). The audio signal was recorded in a PCM format at 16 kHz sampling, 16 bit quantization. The video was stored as AVI files with 24-bit color depth, 30 frame/s.

5. Subjective evaluation

We conducted subjective evaluation experiments to investigate the effect of the various information on the human estimation of the subject's internal state. The information included the system's prompt, the subject's non-verbal behavior and the user's answer utterance.

5.1. Sessions

We split the recorded video and speech into sessions, which included one system prompt and the subject's answer to the prompt. When the subject did not make an answer, we regarded the section from the beginning of the system prompt to just before the next prompt as a session. As a result, we obtained 793 sessions from the recorded video.

5.2. Clips for subjective evaluation

As mentioned above, we prepared four kinds of clips (Clips A, B, C and D) for each of the sessions.

In the majority of the collected 793 sessions, the subjects immediately answered the question, which should be classified into state C (the user had no problem answering the question) [5]. The main interest of this work is how to discriminate the users of state A and B. Therefore, we excluded the sessions where the subject answered the question within 5 s. As a result, we used 255 sessions for the evaluation experiment.

5.3. Evaluation procedure

We employed 18 evaluators (13 males and 5 females) who did not participate in the dialog. We split the evaluators into four groups: Group A, B, C and D, to each of which Clips A, B, C and D were presented, respectively. After watching one session, the evaluator was asked to choose an answer of the following question among the three choices:

Q: How do you evaluate the behavior of the subject with respect to the system's question?
1) The subject did not understand the question (state A).
2) The subject understood the question and took a time to prepare an answer (state B)
3) The subject understood the question and answered it immediately (state C).

6. Experimental result

6.1. Analysis of the result of Group A

First, we investigated the consistency of the judgment made by the evaluators in Group A. We used Cohen's κ to assess the degree of agreement between evaluators. As a result, κ were from 0.46 to 0.55, which showed moderate agreement between the evaluators.

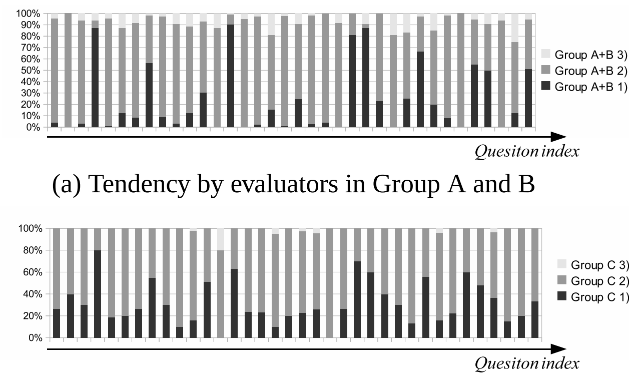
6.2. Comparison between Group A and B

We can assess the effect of the subject's answer to the judgment by comparing the results from Group A and B. We classified each of the sessions by a majority vote for each of Group A and B, and observed agreement between the judgments of two groups. As a result, we obtained $\kappa = 0.59$, which were almost same agreement between evaluators among Group A. This result suggests that effect of the subject's answer on the evaluator's judgment is small.

We also investigated the examples where evaluators of Group A and B gave different judgments. One typical example is that a subject answered "I don't know" after a long silence. In this case, evaluators of Group A tended to regard that example as choice 1 (state A) while those of Group B (who did not hear the answer of the subject) tended to judge it as choice 2 (state B). In addition, the evaluators of Group A seemed to refer the pitch and power of the answer utterance as cues to the subject's confidence of the answer.

6.3. Comparison between Group AB and C

We investigated the effect of the content of the system prompt on the evaluators' judgment by comparing the results of Group C with those of Group A and B. Agreement of the majority vote results of Group B and C was not good ($\kappa = 0.30$), which suggests that the content of the system prompt had some effect on the evaluators' decision.



(a) Tendency by evaluators in Group A and B

(b) Tendency by evaluators in Group C

Figure 1: Tendency of judgment by evaluators

Next, we made a question-by-question analysis to make further analysis for the results. In this analysis, we observe a tendency of judgment for a specific question. Let $n_{qj}(G)$ be the number of evaluations with value $j \in \{1, 2, 3\}$ for question q by the evaluators in Group G. Here, the value of evaluation corresponds to the choices described in section 5.3. Then we calculate

$$\boldsymbol{R}_q(G) = (r_{q1}(G), r_{q2}(G), r_{q3}(G)) \tag{1}$$

$$r_{qj}(G) = \frac{n_{qj}(G)}{\sum_{j' \in \{1,2,3\}} n_{qj'}(G)} \tag{2}$$

$\boldsymbol{R}_q(G)$ is a vector that reflects a tendency of judgment made by evaluators in Group G for question q.

Figure 1 shows the tendency of judgment by evaluators. Figure 1(a) shows $\boldsymbol{R}_q(G_A \cup G_B)$ and Figure 1(b) shows $\boldsymbol{R}_q(G_C)$, where G_A, G_B and G_C are sets of evaluators in Group A, B and C, respectively. Note that, although we prepared 44 questions for the experiment, no responses that took more than 5 s were observed for 8 questions; therefore, we used only 36 questions for the evaluation.

From figure 1, we can see that most questions were judged as either 1 or 2, as intended (most of 3 (state C) had small duration from the prompt to the answer, and thus were excluded from the evaluation). Another observation is that ratio of value 1 for each question had similar tendency for both sets. To confirm this, we investigated the correlation between $r_{q1}(G_A \cup G_B)$ and $r_{q1}(G_C)$ for all questions. Figure 2 shows the scattergram. The correlation coefficient between $r_{q1}(G_A \cup G_B)$ and $r_{q1}(G_C)$ is 0.77, which shows that judgments by the evaluators of Group A+B and C have some similarity.

This observation suggests that tendency of judgment for specific question with and without linguistic information of the response is similar, which justifies our approach to estimate the user's internal state using nonverbal information [5]. However, it is also true that the

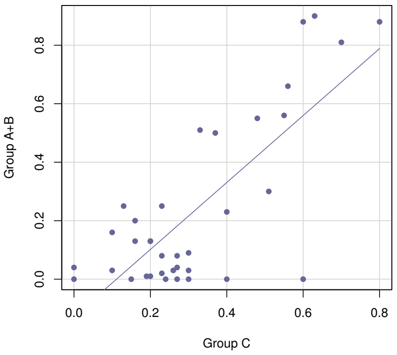


Figure 2: Scatter plot of ratio of value 1

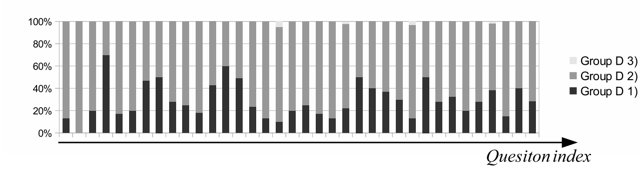


Figure 3: Tendency of judgment by Group D

linguistic information of the response affects the judgment. In the judgment of Group C, the ratio of value 1 is bigger than that of Group A+B (average value of $r_{q1}(G_A \cup G_B)$ is 0.24 and that of $r_{q1}(G_C)$ is 0.32), which seems that the judgment became more difficult without observing the subjects' response, and thus the judgments by Group C were more random than by Group A and B.

6.4. Comparison between Group C and D

In this section, we analyze the effect of audio information on the decision of the internal state by comparing evaluation results by evaluators of Group C and D.

First, we calculated the agreement of the majority vote of Group C and D. The result was $\kappa = 0.41$, which showed moderate agreement. Compared with the evaluations by Group C, that of Group D tended to be evaluated as 2 (state B), as shown in Figure 3. One observation was that there were users' responses in which the users moved their lips but said nothing. When speech was missing, the evaluators tended to judge such responses as value 2 (thinking), while the evaluation was value 1 (being perplexed) when presented with speech. as value 1 (being perplexed). Other examples were filled pauses and fillers; subjects with filled pauses and fillers were judged as 2 when speech was presented, but the filler had no effect when speech was omitted.

7. Conclusions

In this paper, we focused on the user's internal state after the system's prompt and before the user's first utterance. In our previous work, effect of contents of system prompt and user's utterance was unclear. By analysis presented in this paper, we can conclude the following three observations:

1. The user's answers had a small effect on judgments of the evaluators.
2. Content of the system's prompt utterance had a considerable effect on judgments (probably it became more difficult to judge without the prompt), but the tendency of judgments was still similar.
3. Audio information had also a large effect, but the evaluators still could judge with only the visual information. Audio-visual synchronization (such as lip motion and speech) had an effect on the judgment.

In future work, we will investigate methods to determine the user's internal state automatically in real-time by using audio and visual information, and implement it to the dialog system.

8. Acknowlegment

This work is a part of the project "Experimental challenges for dynamic virtualized networking resource control over an evolved mobile core network.a new approach to reduce massive traffic congestion after a devastating disaster," supported by Ministry of Internal Affairs and Communications, Japan.

9. References

[1] A. Kobsa. User modeling in dialog systems: Potentials and hazards. *AI&Society*, 4:214–231, 1990.

[2] A. N. Pargellis, H.-K. J. Kuo, and C.-H. Lee. An automatic dialogue generation platform for personalized dialogue applications. *Speech Communication*, 42:329–351, 2004.

[3] R. Gajšek, V. Štruc, S. Dobrišek, and F. Mihelič. Emotion recognition using linear transformations in combination with video. In *Proc. Interspeech*, pages 1967–1970, 2009.

[4] 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 *Proc. Conf Human factors in computing systems*, pages 257–264, 2003.

[5] Y. Chiba, M. Ito, and A. Ito. Estimation of user's internal state before the user's first utterance using acoustic features and face orientation. In *Proc. HSI*, 2012.

[6] S. E. Brennan and M. Williams. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *J. Memory and Language*, 34(3):383–398, 1995.

[7] M. Swerts and E. Krahmer. Audiovisual prosody and feeling of knowing. *J. Memory and Language*, 53(1):81–94, 2005.