Spoken Dialogue System Using Recognition of User's Feedback for Rhythmic Dialogue

Shinya Fujie, Riho Miyake & Tetsunori Kobayashi

Department of Computer Science
Waseda University, Japan
fujie@pcl.cs.waseda.ac.jp

Abstract

The recognition method of user's feedback during the system's utterance is proposed and its application to the spoken dialogue system is discussed. In human conversation, we can know the dialogue partner's internal state by receiving such feedbacks. Our research topics are (1) developing the prosodic information based feedback recognizer and (2) appropriately controlling the system's utterance timing along with the user's feedbacks. The implemented recognizer can distinguish between back-channel and ask-back word-independently with prosodic information based features and statistical recognition method. Experiments of the spoken dialogue system with this function reveals when it should generate the next utterance after receiving the user's feedback.

1. Introduction

In human conversation, participants give some feedbacks during the partner's utterance. These feedbacks are essential to realize a natural conversation. By generating and recognizing them properly, we can adjust the rhythm of the conversation and we can know the partner's internal state. Most of conventional spoken dialogue systems, however, fixed the timing of turn-taking rigidly. They assume that the end point of the speech recognition of user's utterance is the transfer of the turn from the user to the system, as well as the end point of the system's speech synthesis is that from the system to the user. This kind of systems has to wait until the end of the user's utterance, once they finish their utterance and transfer the turn to the user. Moreover, users also have to wait for the end of the system utterance. In order to make the conversation between human and system more rhythmical and effective, the system should generate the feedbacks during the user's utterance, and also should recognize the user's feedbacks during its utterance.

Recently, several studies focus on the generation of such feedbacks[1, 2]. We also proposed the spoken dialogue system which can generate the back-channel feedbacks appropriately during the user's utterance[3]. In contrast, in this study we aim at recognizing user's feedbacks during the system's speech and controlling the system's next utterance timing by its result.

In order to recognize user's internal state in their feedbacks, it is necessary to reveal what information brings the state by feedbacks. Nakano et al. developed the system which can recognize user's back-channel feedback by using linguistic information[1]. The changes of user's internal state, however, may appear in the style of the utterance, even though the spoken words are the same. In addition, there are many studies on the correlation between prosodic information and emotions, for example, Lee et al.[4]. In this study, we aim at detecting user's internal state from their feedbacks using prosodic information.

Finally, we reveal how to control the system's utterance with the user's feedback, the result of the feedback recognition, through the experiments with the spoken dialogue system.

2. Feedback Recognition

2.1. Target

The target is to recognize the user's internal state from his/her feedbacks during the system's utterance. We define two internal states, which may affect the dialogue management, based on the user's understanding.

Back-Channel This represents that the user normally listens to the system's utterance and understands what it means. Additionally, this is the sign which means the user wants the system to continue its speech.

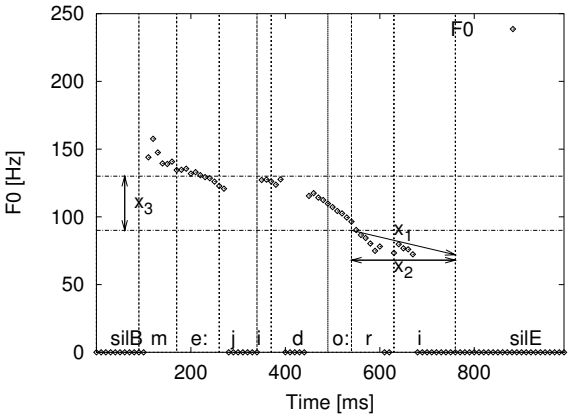
Ask-Back This represents that the user has a trouble with listening to the system's utterance or he/she cannot understand what it means.

The aim of this section is to recognize the user's internal state (back-channel/ask-back) from the prosodic information of the feedback.

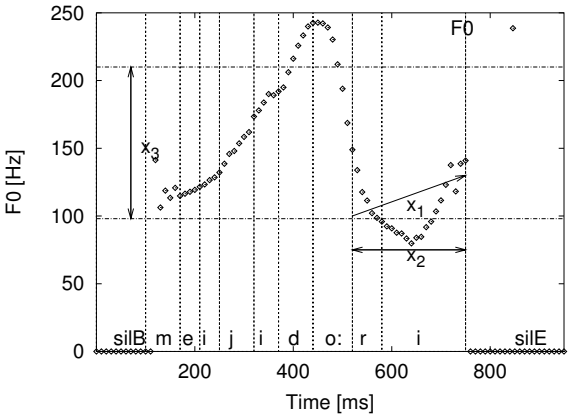
In order to see what expression to be often used as the feedback, we observed human-human conversations. In these observations, two expressions, interjection and repetition, were often used. Interjection means a short word, such as "hai(yes)," "e(what)," etc. Repetition means a fully or partially repeating of the speaker's utterance. Particularly in repetitions, it changes dependently to the style of the utterance that the given feedback is either back-channel or ask-back, even though the linguistic content is the same. Thus, prosodic information is introduced to recognize the feedback.

2.2. Data

In order to collect a large number of feedback data of back-channel and ask-backs, we recorded the users' feedbacks to the utterances synthesized by the spoken dialogue system. We prepared five scenarios for recording of interjections and four for recording of repetitions. Each scenario is a long sentence constructed by several phrases, and each phrase boundary has a short pause for user to produce a feedback. The numbers of these short pauses are 39 for interjections and 20 for repetitions throughout all scenarios. Subjects are 10 male students. They were ordered to produce a feedback in the short pause. For each scenario, data were recorded with four different combinations of state(acknowledgment or ask-back) and expression (instructed or free).



(a) back-channel



(b) ask-back

Figure 1: Examples of prosodic feature extraction for feedback recognition.

2.3. Recognition method

We apply the F_0 extraction and phoneme alignment to the recorded utterances. Difference between back-channel and ask-back appears in the gradient of F_0, so we use that throughout the utterance as the feature value to distinguish them. It is sufficient to use the gradient of F_0, because interjections are very short and composed of a few phonemes. On the other hand, repetitions are longer and consist of more phonemes. As shown in Fig. 1, it is difficult to distinguish repetitions with the same way as the interjection. Thus, following 3-dimensional feature is introduced.

x_1 : the gradient of F_0 of the last mora

x_2 : the duration of the last mora

x_3 : the standard deviation of F_0

3. Experiment

3.1. Recognition experiment

We made two kinds of model sets, (A) and (B).

(A) two models learned individually with feature vectors of interjection and repetition respectively.

(B) one model learned with all feature vectors of interjection and repetition together.

Table 1: The experimental results of model set (A), which contains models learned individually for interjection and repetition respectively.

	back-channel			ask-back		
	correct	total	recog. rate	correct	total	recog. rate
interjection	710	766	92.7%	607	670	90.6%
repetition	382	461	82.9%	259	322	80.4%
total	1092	1227	89.0%	866	992	87.3%

Table 2: The experimental results of model set (B), which contains a single model learned with all feature vectors of interjection and repetition together.

	back-channel			ask-back		
	correct	total	recog. rate	correct	total	recog. rate
interjection	697	756	92.2%	597	658	90.7%
repetition	363	461	78.7%	270	322	83.9%
total	1060	1217	87.1%	867	980	88.5%

In (B), we use 3-dimensional feature not only for repetition but also for interjection. A feedback is recognized as one internal state, back-channel or ask-back, with Bayesian classifier. Gaussian mixture model is calculated with the feature vectors for each internal state. The correct answers are the samples that three subjects listened to and at least two of them agreed.

Table 1, 2 show the recognition results of 10-fold cross validation. The test set consists of all the data of one subject, which is not in the learning set. From these results, there is a small difference in recognition rate between (A) and (B). To apply the recogizer using model (A) to the dialogue system, we must recognize the expression of the feedback is interjection or repetition before the recognition of back-channel and ask-back. If this expression recognition fails, the recognition system performance will deteriorate. Therefore, we can say that (B), one model learned with all feature vectors, has an advantage with the object of application to the spoken dialogue system.

3.2. Criterion for evaluation for application to dialogue system

By considering the recognition result of the feedback, the system can determine its behavior more reasonably. It can continue its utterance when the recognition result is back-channel, while it stops its utterance and switch to another utterance according to user's state when the result is ask-back. Then, if system mistakes back-channel for ask-back, it stops its utterance even when it should keep speaking. Unnecessary pause in speech interrupts smooth communication and causes discomfort for users. Thus, we should reduce the back-channel recognition error. In order to do that, we introduce likelihood ratio threshold τ. When the likelihood ratio of back-channel and ask-back exceeds τ, that is

$$\frac{P(x|\text{back-channel})}{P(x|\text{ask-back})} > \tau \qquad (1)$$

then the result is acknowledgment. Figure 2 shows recognition rate variance with changing of τ. This graph represents that the sacrifice of the recognition rate of ask-back compensates a certain level of the recognition rate of back-channel. For example, the recognition rate is 90.8% in back-channel and 81.1% in ask-back when the number of mixture is 32.

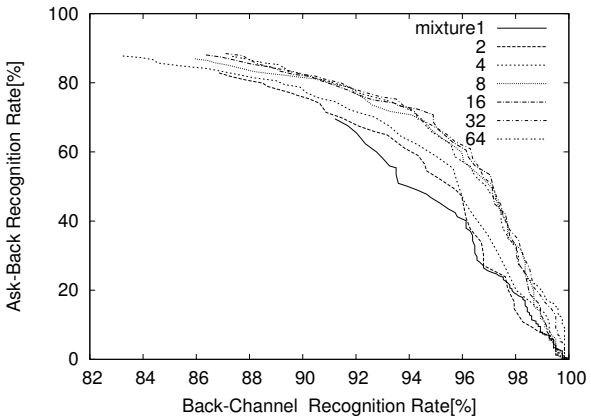


Figure 2: Recognition rate variance for the likelihood ratio τ. By increasing τ, the back-channel recognition rate gets higher while the ask-back recognition rate gets lower. The dialogue system can prevent the unnecessary pause of its utterance by reducing the recognition error of ask-back.

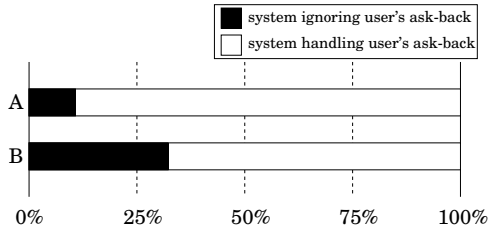


Figure 3: The results of the subjective tests of the spoken dialogue system with the feedback recognition.

3.3. Subjective dialogue experiment

In order to confirm the effects of our methods, we performed subjective tests.

We prepared the systems according to the availability of the feedback generation/recognition. The back-channel feedback generation function is realized with the method described in [3]. Conversations between the user and the system were recorded into a video with different scenario for each pair. Length of each video is about one minute.

After watching these videos, 28 subjects answered which conversation they preferred for each pair. The order of watching is random for each pair. Optionally they also answered the reason why they preferred it.

The evaluation results are shown in Fig. 3. According to the results, the system with the recognition function were preferred in both pairs. The dialogue system is able to speak the next utterance quickly by recognizing the back-channel and give the particular description to the user by recognizing the ask-back while the system itself is speaking. Most of the comments from subjects who preferred the system without the recognition function mention to the back-channel feedback generation function. Probably they couldn't pay attention to the availability of the recognition function. By the fact that the system with the function is overwhelmingly preferred particularly in pair A, we see that the feedback recognition function has an effect to make the impression of the conversation good.

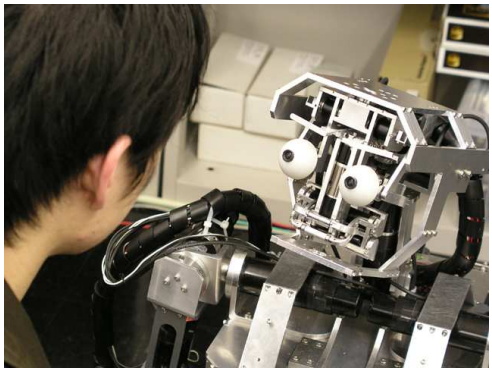


Figure 4: Conversation robot ROBISUKE

4. Application of user's back-channel feedback

The dialogue system described in 3.3 can handles the user's ask-back in its utterance by changing the previously planned utterance. For the back-channel feedback, it simply generates the planned utterance immediately after it receives a feedback. Previous studies treated the user's response in the system utterance as the interruption of it. However, back-channel feedbacks are also very important to realize rhythmic conversation between human and system.

In this section, we investigate how the user's back-channel feedback timing changes as the system utterance changes, as well as how the user's impression of the system utterance changes as the system utterance timing for the user's back-channel feedback changes.

4.1. Experiment setup

Subjects, playing the role of user, were ordered to give the back-channel feedback in synchronization with the system utterance. The system is implemented on the conversation robot ROBISUKE, shown in Fig.4. The system was configured with the following parameters for each session.

Rate of Speech (S) System speaks in the rate of either *Fast*, *Normal*, or *Slow*, in each session. *Fast* is 1.1 times *Normal* in rate, while *Slow* is 0.9 times *Normal*.

Start Timing of System Utterance (Δt^S) System starts its remaining utterance either 0ms, 300ms, or 600ms after the user's back-channel feedback in each session. In order to see the case which system ignores the feedback, we prepare the system which starts its remaining utterance either 300ms or 600ms after the end of its previous speech.

Start Timing of Controlling System Gaze (Δt_s^E) It may encourage the user's utterance or feedback that system casts the eyes on the user when it finishes its speech. System starts the gaze control either 300ms before, 300ms after, or immediately as the end of utterance. We also prepare the system which is always staring at user ($always$) and never watches user ($never$).

End Timing of Controlling System Gaze (Δt_e^E) System finishes casting the eyes on the user either 300ms before, 300ms after, or immediately as the start of utterance.

In the combination of the above-mentioned parameters, the system speaks utterances under the scenario randomly selected from five prepared scenarios. Each scenario is the simple ex-

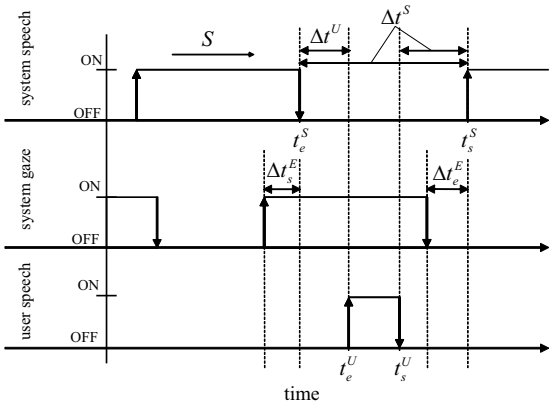


Figure 5: The parameters for controlling of the system utterance timing against the user's back-channel feedback.

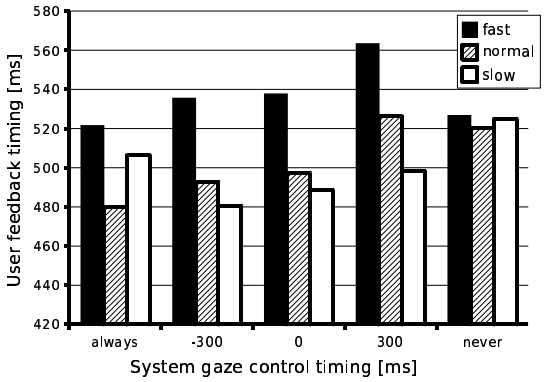


Figure 6: The changes of the user's feedback timing according to the system's utterance speed and the timing of the gaze controlling.

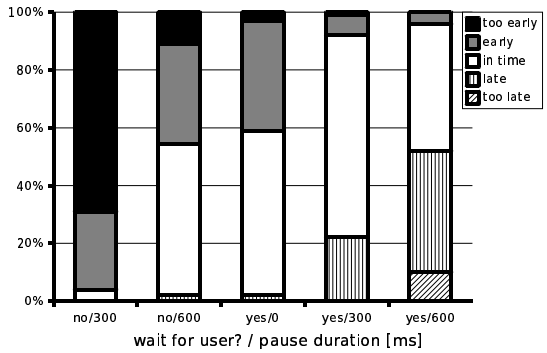


Figure 7: The results of subjective evaluations for the changes of the system's utterance timing.

planation of a restaurant, and has three pauses for the user to give back-channel feedback.

There are 144 combinations of the parameters. The sessions of all parameter combinations are executed for each subject. For each session, the length of the pause between the end of system utterance and the subject's back-channel feedback is recorded. At the same time, subjects answer the question, "How is the timing of the robot utterance after your feedback?" by choosing the prepared answers, such as "too late," "late," "in time," "early," and "too early" after each session. Subjects are 10 male students.

4.2. User's back-channel feedback timing

Figure 6 shows the changes of the user's back-channel feedback timing. This graph indicates:

- User's back-channel feedback timing becomes earlier as the system's utterance speed becomes slower.
- User's back-channel feedback timing becomes earlier as the timing of system's gaze control becomes earlier.

The former result shows that the slower utterance is easier for subjects to estimate the end point, where they should give the feedback, than the faster one. The latter result shows that the subjects can estimate the end point more easily by observing the motion of the robot gaze around the point.

4.3. Subjective evaluation of the system utterance timing

Figure 7 shows the results of the subjective evaluation. This graph shows that the utterances that the system gave 300ms after the user's feedback end are preferred. Moreover, the utterances given at the same time as (0ms after) the user's feedback end are thought to be early, while the utterances given 600ms after the user's feedback end are thought to be late. Therefore, when the user give some back-channel feedback during the pause between the system's utterances, it should begin the next utterance 300ms after the end of the feedback.

5. Conclusion

In this paper, we proposed and implemented back-channel feedback recognition methods for more natural conversation on spoken dialogue system. We used prosodic information and achieved the high recognition rate enough to apply the results for the spoken dialogue system. Through the subjective tests, we confirmed the effectiveness these function and revealed how to control the system's utterance timing against the user's feedback.

We aim at the improvement of our methods. The linguistic information in the user's utterance brings important information for recognition, especially in the repetition. Considering these sorts of useful information is our next target. Our future work also includes the confirmation of the effectiveness of our system from the point of view by the efficiency of the dialogue, such as task achievement, and so on.

6. References

[1] Nakano, M.; Dohsaka, K.; Miyazaki, N.; Hirasawa, J.; Tamoto, T.; Kawamori, M.; Sugiyama, A.; Kawabata, T., 1999. Handling rich turn-taking in spoken dialogue systems. *6th European Conf. on Speech Communication and Technology, Eurospeech '99*, 1167-1170.

[2] Kitaoka, N.; Takeuchi, M.; Nishimura, R.; Nakagawa, S., 2005. Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems. *Journal of The Japanese Society for Artificial Intellignece*, 20(3), SP-E, 220-228.

[3] Fujie, S.; Fukushima, K.; Kobayashi, T., 2005. Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. *9th European Conf. on Speech Communication and Technology, Interspeech2005*, Portugal, 889-892.

[4] Lee, C.M.; Narayanan, S.S., 2005. Toward detecting emotions in spoken dialogs, *IEEE Trans. Speech and Audio Processing*, 13(2) 293-303.