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

3.1. Recognition experiment

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.

<table>
<thead>
<tr>
<th></th>
<th>back-channel</th>
<th>ask-back</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>total</td>
<td>recog. rate</td>
</tr>
<tr>
<td>interjection</td>
<td>710</td>
<td>766</td>
</tr>
<tr>
<td>repetition</td>
<td>382</td>
<td>461</td>
</tr>
<tr>
<td>total</td>
<td>1092</td>
<td>1227</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>back-channel</th>
<th>ask-back</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>total</td>
<td>recog. rate</td>
</tr>
<tr>
<td>interjection</td>
<td>697</td>
<td>756</td>
</tr>
<tr>
<td>repetition</td>
<td>363</td>
<td>461</td>
</tr>
<tr>
<td>total</td>
<td>1060</td>
<td>1217</td>
</tr>
</tbody>
</table>

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.

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 \( \tau \). When the likelihood ratio of back-channel and ask-back exceeds \( \tau \), that is

\[
P(x|\text{back-channel}) > \frac{P(x|\text{ask-back})}{\tau}
\]

then the result is acknowledgment. Figure 2 shows recognition rate variance with changing of \( \tau \). 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.
4. Application of user’s back-channel feedback

The dialogue system described in 3.3 can handle 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 (ΔtE₁) 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 (ΔtE₂) 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 (ΔtE₃) 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-
We confirmed the effectiveness of these functions and revealed how they can improve the spoken dialogue system. Through the subjective tests, we used prosodic information and its application to spoken dialogue systems. 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.

5. Conclusion

In this paper, we proposed and implemented back-channel feedback recognition methods for more natural conversation on spoken dialogue systems. 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 of these functions 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


