2nd International Conference on Autonomous Robots and Agents
December 13-15, 2004 Palmerston North, New Zealand

A Conversation Robot with Back-channel Feedback Function based on Linguistic and Nonlinguistic Information

Shinya Fujie, Kenta Fukushima, Tetsunori Kobayashi
School of Science and Engineering,
Waseda University, Tokyo, Japan

Abstract

A conversation robot which can generate a back-channel feedback appropriately to the user is developed. The contents of these feedbacks depend on the contents of the dialogue partner's utterance, which are provided by the speech recognition. In order to determine the content of the feedback earlier than the end of the utterance, we used finite state transducer based speech recognizer which outputs the content of the feedback. And we used prosody information, especially the fundamental frequency(F0) and the power of the utterance, to extract the proper timing of the feedback. We implemented these modules and applied them to the spoken dialogue system on the humanoid robot ROBISUKE. Experimental results show the effectiveness of our methods.

Keywords : Conversation robot, back-channel feedback, spoken dialogue system, prosody, FST

1 Introduction

1.1 Back-channel Feedback in Conversation

In human conversation, we exchange the turn of talking one another. The natural exchange of turn realizes the natural conversation. Most of conventional spoken dialogue systems 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. And also they assume the end point of the system's speech synthesis is that from the system to the user. Those systems have to wait until the end of the user's utterance, once they finish their utterance and transfer the turn to the user. However, in human conversation, while the speaker speaks something, the listener feeds back his/her state in some ways. For example, the listener nods to show he/she listens to the speaker carefully, and the listener repeats the speaker's words to feed back how he/she understands what the speaker says. By receiving these feedbacks, the speaker feels easy to talk, because he/she is able to knows the listener listens to the utterance and how the listener understands. These feedbacks are generated by the listener unconsciously. It is called "back-channel feedback." If a spoken dialogue system can generate the back-channel feedback appropriately, the conversation becomes more effective and natural.

1.2 Studies about Back-channel Feedback Generation

To generate the back-channel feedback appropriately, the dialogue system must decide the content and the timing of the feedback according to the user's utterance. There are several studies about the back-channel feedback generation. They can be roughly classified into ones that use linguistic information and ones that use non-linguistic information.

Nakano et al. developed the spoken dialogue system which can do natural turn-taking by the incremental understanding using the context[1]. This system detects the speech recognition result earlier than the end of the utterance, and executes the language and semantic processing incrementally. The system decides what should be asked to the user and it synthesizes the content as the feedback. This kind of system can generate the feedback with the appropriate content, but it is afraid to make the user uncomfortable because it doesn't control the timing of the feedback.

Ward et al. developed the system that generates the back-channel feedback based on the rule that a feedback is generated after the utterance which contains the low pitch for a certain period of time[2]. Okato et al. detected the timing of the feedback using the prosodic templates obtained from the spoken dialogue corpus, and developed the feedback generation system with the same method[3]. Koiso et al. proposed the model that predicts the end/succession of the utterance by the analysis of correlation among the word, pitch pattern and energy pattern in the ending part of the utterances[4]. The system developed by Takeuchi et al. controlled the timing of the feedback generation, using the prosodic information and the part of speech of the spoken word[5]. This kind of system can decide the timing of the feedback generation appropriately. However, it cannot decide the contents of the feedback.

1.3 Proposed Method

In this paper, we aim at the system that can generate the back-channel feedback with the appropriate content at

the appropriate timing. As described above, the content of the feedback depends on linguistic information, i.e. the content of the utterance, and the timing depends on the style of the utterance.

In order to respond with the appropriate content in the middle of the user's utterance, the system must detect the results of the speech recognition before the end of the utterance. It seems to be difficult to realize the early detection using conventional speech recognizer, because of the complexity of the data structure and the implementation. In this study, we use the Finite State Transducer(FST) based speech recognizer, which becomes popular recently, to decide the content of the feedback. FST itself has an early detection function, and it can generates any symbols as output by composing several transducers.

The timing detection is realized by processing prosodic information for every frame. In this study, as prosodic information, we use the fundamental frequency (F0) and the logarithmic power of the user's utterance.

Using these two outputs, we implement the spoken dialogue system that can generate the back-channel feedback with appropriate content at the appropriate timing, on the humanoid robot.

2 Early Detection of Feedback Contents

2.1 Speech Recognizer Using FST

FST is an automaton with output symbols. A transducer with expected input/output symbols can be obtained by composing several transducers. The early decision of output symbols corresponding to the given input symbols can be realized by the optimizing the transducers with several operations such as minimization, determinization, and so forth.

In order to recognize user's speech with a large vocabulary continuous speech recognizer(decoder), we have to give it three databases, so called acoustic model, word dictionary and language model. Acoustic model is a probabilistic model that gives an acoustic feature output probability of each phoneme. Word dictionary is a list of phoneme sequences of all words. Language model is a probabilistic model that represents which word sequence is more plausible in a target language. Common decoders use Hidden Markov Model(HMM) for acoustic model and *n*-gram model (bi-gram or tri-gram) for language model.

HMM, word dictionary and *n*-gram language model can be represented by a network, and the recognition (decoding) is done by transition on the network. A single FST with early detection function can be obtained by constructing FSTs of these models individually, composing them into single FST and optimizing it. This kind of network is the most optimized network

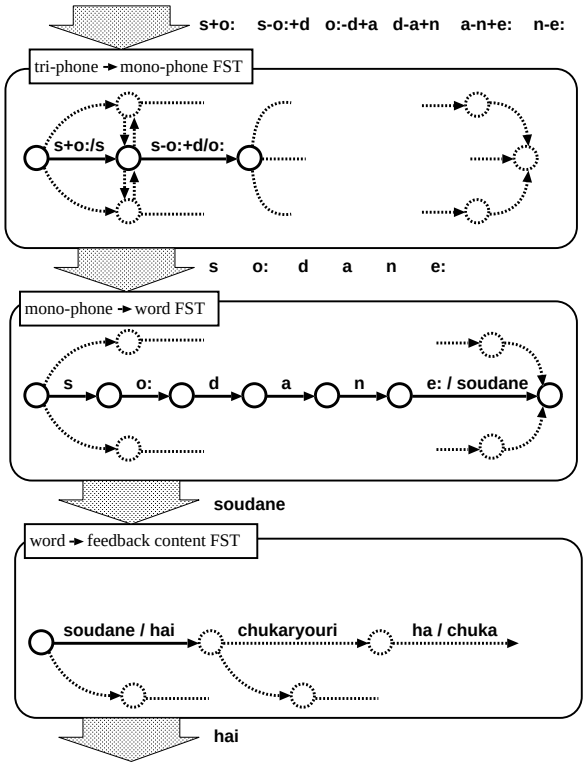

Figure 1: FST with the feedback contents as output symbols.

that can be obtained before the system works. And because the network is the only single one, we can use simple search algorithm, and the decoder implementation is very simple. Thus, recently, many researches focus on the application of FST to the speech recognition[6][7].

2.2 Early Detection of Back-channel Feedback Contents

Generally, the final output of speech recognition is a word sequence. In this study, the final output is the content of the back-channel feedback corresponding to the utterance of the user. We aim to detect it earlier than the end of the utterance using the early detection function of FST.

As shown in figure 1, a single network which accepts the tri-phone as input symbol and outputs the recognized words and the content of the back-channel feedback, is obtained by composing 3 networks. The first network accepts tri-phone as input and outputs mono-phone. The second one accepts mono-phone as input and outputs the recognized words. The third one accepts a word as input and outputs the given word and the contents of the back-channel feedback. By decoding with this network, the decoder outputs the contents of the back-channel feedback on the point where it should generate the feedback. And also, we give the special output symbol `delete` on the point where the feedback is not appropriate any more. Thus, the

feedback that must not be generated can be eliminated by decoding with this network.

We use tools provided by AT&T[8][9] in order to compose and optimize the FST.

3 Back-channel Feedback Timing Detection Using Prosodic Information

3.1 Back-channel Feedback and Prosody

The output of the decoder using FST, described in the previous section, is the appropriate content of the feedback when it should be generated. According to the early detection function of the FST decoder, the output can be obtained before the timing which the system should actually generate the feedback. But the decoder cannot determine the timing.

The decision whether the system should generate the feedback or not depends on the style of the user's utterance rather than the content of it. For example, in the case that the user's utterance is "karakute oishii mono ga tabetai na(I would like have something hot and tasty)." When the user says the part "karakute" with stretching the last phoneme 'e', the system should generate the feedback like "hai(yes)" or the repetition of the word "karai(hot)." On the other hand, when the user says "karakute oishii" at a breath, the feedback is not only unnatural in the conversation but also uncomfortable for the user.

In this section, we aim to detect the appropriate timing that the system should generate the feedback using the F0 and energy of the user's utterance.

3.2 Analysis of Dialogue Data

At first, in order to collect the partner's utterances right before the back-channel feedback, we recorded the simulated conversation between two people. 9 conversations, free topic, 5 minutes long each, were recorded with video camera. The all back-channel feedback parts over the recorded movie are annotated by hand with annotation tool Anvil[10].

We extracted and observed the partner's utterance right before the back-channel feedback. An example of the utterance "ah ichinichi me de (ah, on the first day)" is shown in figure 2. The graph shows the extracted F0 and logarithm power of the utterance. According to the observation of many samples, characteristic appears in the part around 100msec to 500msec before the feedback.

3.3 Feature and Model

In order to detect the proper timing of the feedback, we introduce the pattern recognition technique. The

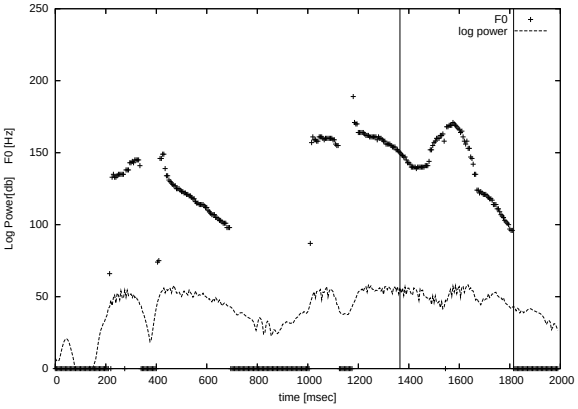

Figure 2: An example of F0 and power extraction of the utterance.

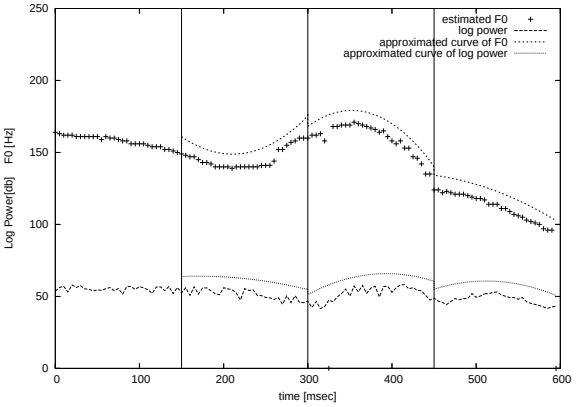

Figure 3: An example of approximated curve of extracted F0 and power.

feature vector consists of the coefficients of the approximated curve of the F0 and logarithm power over the divided 3 regions of the past 450msec region, and the average of the errors between curve and the extracted values. The approximated curve is calculated with least-square method.

$$y'(t) = at^2 + bt + c \tag{1}$$

We adopt the coefficients a and b as the feature. The average of the errors is calculated as follows.

$$\frac{1}{N}\sum_i \left| y_i \quad y'(t_i) \right| \tag{2}$$

Where, y_i represents extracted F0 or logarithm power, and N is the number of the extracted samples. In this study, the sampling rate is 16kHz and the frame shift is 80 samples, so the N is 30 for each region (the duration of one region is 150msec). Three kinds of features about F0 and power are calculated for each region, so the total dimension of a feature vector is 18. Figure 3 shows an example of approximated curve of F0 and power, calculated over the region shown in figure 2.

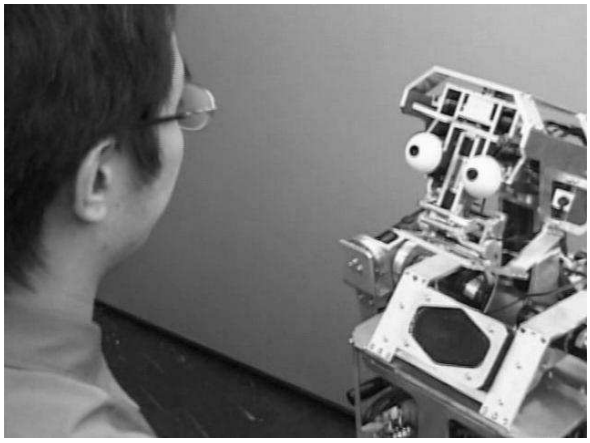

Figure 4: Conversation robot ROBISUKE

The feature vectors were calculated with 119 utterance samples which were obtained from the recorded database described in the previous section. And a Gaussian distribution was learned with these feature vectors. In the recognition stage, the feature vector is calculated, and the output likelihood of it from the learned distribution is evaluated for every frame. When the likelihood exceeds a threshold, it is decided to be the timing for the system to generate the back-channel feedback. The threshold is adjusted by hand before the system works.

4 Back-channel Feedback Function on a Conversation Robot

4.1 System Overview

We applied the proposed FST decoder and prosody process to the spoken dialogue system which we have developed. The spoken dialogue system has been implemented on the humanoid robot ROBISUKE[11] shown in figure 4. The system architecture is shown in figure 5.

The speech of the user's utterance is input into the prosody process module which calculates the feature vector and evaluates the back-channel feedback generation likelihood which we proposed in the previous section, as well as input into the MFCC extraction module which calculates the MFCC which is needed for the speech recognition. In the prosody process module, the sampling rate of the speech is 16kHz, the size of one frame is 1024 samples, and the frame shift is 80 samples(5msec). For the F0 extraction, we use the method proposed by Goto et al[12] which using instantaneous frequency and a comb filter. By more popular method proposed by Talkin[13] (the implementation provided in *ESPS/waves+* and *wavesurfer*) we are able to obtain more precise F0 extraction results. However, Talkin's method use the Dynamic Programming to generate precise results, so for our system, which needs the real-time response,

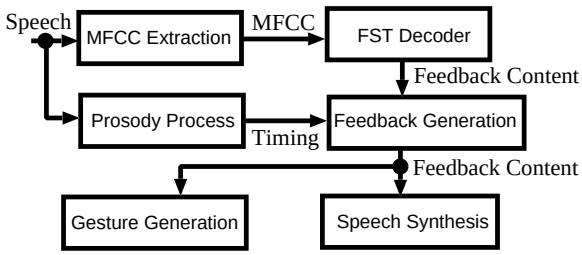

Figure 5: The system architecture

Goto's method is better. In the prosody process module, furthermore, the feature vector described in section 3.3 is calculated for each frame, and the likelihood of the timing for the back-channel feedback is evaluated. As soon as the likelihood exceeds the threshold, it sends a signal to back-channel feedback generation module.

FST decoder receives the MFCC feature and decodes with it. Because a decoding makes progress by spreading the hypotheses, the final output cannot be obtained until the end of the utterance, even if the FST has the early detection function. In our system, therefore, when the most plausible output doesn't change for several frames, it is decided to be the early detected output temporally. If a different output is more plausible than the previous output afterward, it outputs the special symbol `delete` which means the cancel of the previous one and outputs the newly detected one.

As the back-channel feedback generation module receives the content of the feedback from the FST decoder, it stores the content into the buffer. Afterward, as it receives a signal from the prosody process module, it sends the content in the buffer to the speech synthesis module and the gesture generation module.

The speech synthesis module synthesizes the received content. The gesture generation module makes ROBISUKE nod for a short time. This kind of head gesture often occurs along with the back-channel feedback.

5 Experiment

5.1 Target

In order to evaluate the developed system, we perform subjective evaluations. The target of this experiment is to confirm the effects of introducing the FST decoder and the prosody process module.

One of the expected effects of introducing the FST decoder is to enable the system to change the content of the feedback by what the user says. Moreover, because the decoder doesn't generate the output until the user says something for which the system should generate

Table 1: The result of pettern I

A is better	A is slightly better	even	B is slightly better	B is better
2	6	3	8	1

Table 2: The result of pettern II

C is better	C is slightly better	even	D is slightly better	D is better
3	4	0	15	10

a feedback, unnecessary feedbacks are expected to be reduced more than the case of using only the prosody process module.

The prosody process module is expected to enable the system to generate a feedback on the more appropriate timing than the case of using only the FST decoder.

5.2 Experimental Setup

Subjects talk with a pairs of the differently tuned dialogue systems, and compare these systems. The scenario of the conversation is previously prepared. And the evaluation is done by answering the question, "which system do you prefer?" with 5-point scale. We prepared 2 patterns (I and II) of pairs, a subject evaluated twice a pattern. There were 10 subjects for pattern I, and 16 subjects for pattern II.

Pattern I consists of system A which generates the feedback using only the prosody process module, and system B which generates the feedback using both the prosody process module and FST decoder. When the prosody process module sends a signal, system A generates the feedback regardless of the content of the user's utterance. In this experiment, we can see the effect of introducing the FST decoder.

Pattern II consists of system C which generates the feedback using only the FST decoder, and system D which generates using both modules like system B. In this experiment, we can see the effect of the prosody process module.

5.3 Results

The results of two experiments are shown in table 1 and 2. As we can see in table 1, there is few difference in pattern I. On the other hand, in table 2, there is clear difference where subjects prefer the system D which generates depending on both modules, in pattern II.

6 Discussion

The importance of the timing control based on the prosodic information is confirmed from the result of pattern II. On the other hand, the result of pattern I is separated to both sides, so the effects of using linguistic information cannot be seen clearly.

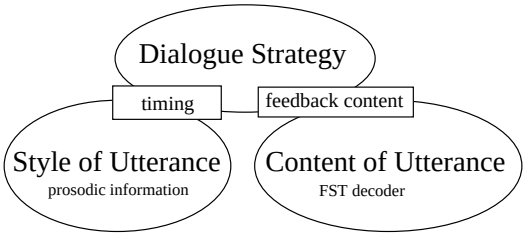

Figure 6: 3 factors to decide the back-channel generation.

Here, we discuss the factors that determine the timing and the content of the back-channel feedback.

The decision whether the back-channel feedback should be generated or not and the content of it seem to depend on the content and the style of the partner's utterance. In actual human conversation, however, they depend on the dialogue strategy of the participant that will generate the feedback.

For example, when the partner says "oishiii mono ga tabetai na (I would like to have something tasty)," the feedback plays a role quite differently depending on its content, "hai(yes)" or "oishii(tasty)." If it is "hai(yes)," it just represents the listener listens to the partner's utterance carefully, and it encourages the partner to continue the utterance. If it is "oishii," it represents the listener understands the partner's word "oishii" correctly in addition to listening carefully, but it doesn't encourage the partner to continue the utterance as strongly as "hai." And basically the feedback "hai" encourages the partner to continue the utterance, but too many feedbacks (even if its content is the same "hai") forces the partner to finish the utterance.

As shown in figure 6, the content and the timing of the back-channel feedback depend on not only the style and the content of the partner's utterance but also the dialogue strategy of the system itself. Therefore, in order to construct the spoken dialogue system that can generate appropriate back-channel feedbacks, we must consider the dialogue strategy of the system too.

7 Conclusion

In this paper, we proposed and implemented the FST decoder that generates the appropriate content of the back-channel feedback, and the method that detects the appropriate timing of the feedback using prosodic information. And we implemented the spoken dialogue system that can generate the back-channel feedback using proposed methods on the humanoid robot ROBISUKE.

Experimental results show the effectiveness of the FST decoder and the prosody process module. It shows that the timing is more important.

Future work includes the considering of the dialogue strategy of the system in order to make the content of the feedback more effective and more natural.

References

[1] M. Nakano, K. Dohsaka, N. Miyazaki, J. Hirasawa, M. Tamoto, M. Kawamori, A. Sugiyama, and T. Kawabata. Handling rich turn-taking in spoken dialogue systems. In *Proceedings of Eurospeech '99*, pages 1167–1170, 1999.

[2] N. Ward. Prosodic features which cue back-channel responses in English and Japanese. *Pragmatics*, 32:1177–1207, 2000.

[3] Y. Okato, K. Kato, M. Yamamoto, and S. Itahashi. Giving 'aizuchi' using prosodic information. *ISPJ Journal*, 40(3):469–478, 1998. (in Japanese).

[4] 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 dialogues. *Language and Speech*, 41(3-4):294–317, 1998.

[5] M. Takeuchi, N. Kitaoka, and S. Nakagawa. Timing detection for realtime dialog systems using prosodic and linguistic information. In *Proceedings of the International Conference Speech Prosody (SP2004)*, pages 529–532, 2004.

[6] Takaaki Hori, Chiori Hori, and Yasuhiro Minami. Speech summarization using weighted finite-state transducers. In *Proceedings of Eurospeech 2003*, pages 2817–2820, 2003.

[7] S. Kanthak, H. Ney, M. Riley, and M. Mohri. A comparison of two LVR search optimization techniques. In *Proceedings of ICSLP 2002*, pages 1309–1312, 2000.

[8] M. Mohri, F. Pereira, and M. Riley. Weighted finite-state transducers in speech recognition. In *Proceedings of ISCA ITRW, ASR 2000*, pages 97–106, 2000.

[9] AT&T Finite State Machine Library. Home page. http://www.research.att.com/sw/tools/fsm/.

[10] M. Mipp. Anvil - a generic annotation tool for multimodal dialogue. In *Proceedings of Eurospeech 2001*, pages 1367–1370, 2001.

[11] Shinya Fujie, Yasushi Ejiri, Hideaki Kikuchi, and Tetsunori Kobayashi. Recognition of paralinguistic information and its application to spoken dialogue system. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2003)*, pages 231–236, December 2003.

[12] Masataka Goto, Katunobu Itou, and Satoru Hayamizu. A real-time filled pause detection system for spontaneous speech recognition. In *Proceedings of Eurospeech '99*, pages 227–230, September 1999.

[13] D. Talkin. *Speech Coding and Synthesis*, chapter 14, pages 495–510. Elsevier Science, 1995.