PROSODY-BASED DETECTION OF THE CONTEXT OF BACKCHANNEL RESPONSES

Hiroaki Noguchi and Yasuharu Den

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayamacho, Ikoma, Nara, 630-0101 Japan
e-mail: {hiroak-n,den}@is.aist-nara.ac.jp

ABSTRACT

Current spoken dialogue systems lack positive feedback such as backchannels, which are common in human-human conversations. To develop more natural human-computer interfaces, the investigation of backchannel-responses are indispensable. In this paper, we propose a method for detecting the precise timing for backchannel responses in Japanese and aim at incorporating such method in future spoken dialogue systems. The proposed method is based on machine learning technique with a variety of prosodic features. It is shown to be effective in automatically deriving rules for detecting the contexts of backchannels. The performance of our method is considerably better than previous methods.

1. INTRODUCTION

Many researchers have reported that people hesitate to talk with spoken dialogue systems due to the lack of positive feedback from the systems such as backchannels, which are common in human-human conversations [3, 6]. To develop more natural human-computer interfaces, the investigation of backchannel-response mechanisms are indispensable. In this paper, we propose a method for detecting the precise timing for backchannel responses in Japanese and aim at incorporating such method in future spoken dialogue systems.

In the proposed method, the contexts for backchannels are detected by using only prosodic features such as fundamental frequency and energy, which are relatively easy to handle by current speech technology. In contrast to the existing methods, which use very limited number of features and hand-made heuristics, we employ a machine learning method with a variety of prosodic features which might be relevant to the detection of the backchannel context. It will be shown that our method is effective in automatically deriving rules for detecting the contexts of backchannels and that it performs considerably better than previous methods.

In Section 2, we review related works on backchannels in Japanese conversation and automatic detection of the timing for backchannels. In Section 3, we describe the spoken dialogue corpus used in our study and provide our definition of backchannels. In Section 4, we conduct a psychological experiment in order to categorize positive and negative contexts for backchannels which are common to average humans. In Section 5, we obtain, by using decision tree learning method, prosodic cues which best discriminate the positive and negative contexts for backchannels. In Section 6, we summarize the paper.

2. RELATED RESEARCH

Backchannels are short utterances, such as `uh-huh' and `yeah' in English and `hai' and `ee' in Japanese, produced by a hearer during a speaker's speech. They are produced not at random but at their appropriate timing. Many researchers have tried to speculate about the factors which determine the timing of backchannels.

Maynard [5] mentioned that in Japanese conversations, a speaker often provides cues for inducing backchannels from a hearer at or around the ends of pause-bounded phrases. She suggested as lexical cues sentence final and interjectory particles, e.g., `ne' in Japanese, followed by a pause. Other researchers also suggested prosodic cues such as rise-fall intonation [2].

Several speech engineers have been working on prosody-based detection of the context of backchannels. Ward [8] reported a heuristics which states that a hearer should respond with a backchannel when a low pitch region in a speaker's speech lasted longer than 150 msec. With this heuristics, he achieved a recall of 53% and a precision of 33%. Okato et al. [6] reported that there are specific pitch patterns in the region of 200 to 400 msec before backchannels. They simulated backchannel responses by a template-matching technique, achieving a recall of 77% and a precision of 33%.

Following Maynard [5] and Okato et al.'s [6] observations, we utilize prosodic features around the ends of pause-bounded phrases.

3. CORPUS

The spoken dialogue corpus used in this study was collected at Nara Institute of Science and Technology (hereafter, NAIST) under the following conditions:

- face-to-face dyadic conversations
- free talk on the topic chosen by the subjects from among the pre-determined list
- recorded in a soundproof room
- no partition between the speakers
- using headset type microphones (but without headphones)
- separate channel for each speaker and sampled in high quality at a rate of 20kHz

We transcribed total of 40 minutes dialogues by 3 different pairs of subjects. Speech materials were divided into pause-bounded phrases delimited by pauses longer than 100 msec, yielding 1875 such phrases.

Original (Japanese)	Translation (English)
L: ya mata gottui / sugoi sensyuyan	he would be a very / great player, wouldn't he?
R: utuno	is he a good hitter?
L: ya sodatikatatokani yotte wakarankedo- / kankyootokani yotte-	well, it depends on his breeding / on his surroundings
R: **aa**	**uh-huh**
L: kedo sondakeno sainoowa arutte yuunga- aruyan	but he would have a talent for that, don't you think so?

Figure 1: Excerpt from the corpus with translation into English on the right column. `L' and `R' identify speakers. Each line corresponds to a conversational move and a `/' indicates a boundary between phrases. Backchannels are in boldface.

We labeled backchannels in the corpus based on their forms and functions. Expressions such as `hai,' `ee,' and `un' in Japanese were judged to be backchannels unless they constituted conversational moves such as an answer to a yes-no question [1]. We found total of 144 backchannels in the corpus.

Figure 1 shows an excerpt from the corpus with an example of backchannels.

4. CATEGORIZING CONTEXTS FOR BACKCHANNELS

4.1. Goal

As mentioned in Section 1, we aim at identifying the features of contexts for backchannels using decision tree learning method. This method requires training data of both negative and positive instances. When collecting these training data from the spoken dialogue corpus, the following problems arise:

1. It is not appropriate to consider as positive cases *only* those contexts where backchannels are found in the corpus.
2. It is not appropriate to consider as positive cases *all* those contexts where backchannels are found in the corpus.
3. It is necessary to consider as negative cases *other* contexts *than* those where backchannels are not found in the corpus.

Problems 1 and 2 Backchannels are considered as optional responses [5]; in fact, the frequency of backchannels found in our corpus varies from speaker to speaker. One person may not respond with a backchannel at the place where another person does, and vice versa. Thus, the spoken dialogue corpus may not contain all backchannels that might have occurred, or it may contain some backchannels that should not have occurred. Using only and all those contexts identified in the corpus as positive cases would result in very low recall and low precision.

Problem 3 Decision tree learning method needs not only positive cases, but also negative cases. One simple way to make negative cases is to collect cases in the corpus where backchannels did not appear. This, however, is not adequate because of the optionality of backchannels; one person may respond when another person may not. Therefore, using these negative cases for learning would result in very low precision.

Considering the above, we do not directly use the contexts found in the corpus. Rather, we identify the contexts where people *commonly* respond with backchannels and *commonly* do not. To do this, we conducted a psychological experiment in which subjects were requested to respond, by hitting keys, to the speech materials at the timing of backchannels.

4.2. Method

Subjects 18 graduate students of NAIST, all native speakers of Japanese (9 males and 9 females).

Materials We selected, from the corpus, 176 stimuli, each of which consists of several pause-bounded phrases and constitutes a single conversational move [1]. We excluded those cases that had difficulty in understanding or listening, was too short to respond, or contained only one pause-bounded phrase. Note that, by our definition of backchannels, only responses to move-internal pause-bounded phrases are considered to be backchannels.

The average number of pause-bounded phrases contained in a stimulus was 2.91 (= 512/176).

Procedure The subject was asked to respond by hitting the space bar whenever he or she thought it appropriate to respond with a backchannel while listening to a stimulus. Each pause-bounded phrase within a stimulus was judged to be followed by a backchannel if the subject had responded within 500 msec after the end of that phrase.

Each subject was given 88 stimuli at random order without discourse contexts. Each stimulus was used for 9 subjects.

4.3. Results

For each pause-bounded phrase, we counted the number of subjects who responded to that phrase. Then, we also classified the phrases according to the number of the responding subjects.

Figure 2 shows the histogram of the phrases classified by the number of the responding subjects. In order to clarify the deviation from the random distribution, the figure shows the difference from the binomial distribution with a mean of 3.1, which is the average number of subjects responding to one phrase.

From Figure 2, we can see the following:

1. There are more phrases than expected to which none or only one subject responded. (Group A)
2. There are more phrases than expected to which more than 5 subjects responded. (Group C)
3. There are less phrases than expected to which 2 to 5 subjects responded. (Group B)

Thus, we can conclude that there are both contexts where people commonly respond with backchannels, namely, *positive* contexts, and those where people commonly do not, namely, *negative* contexts. We obtained 106 cases for the positive contexts and 98 cases for the negative contexts, leaving 132 cases as unclassified,

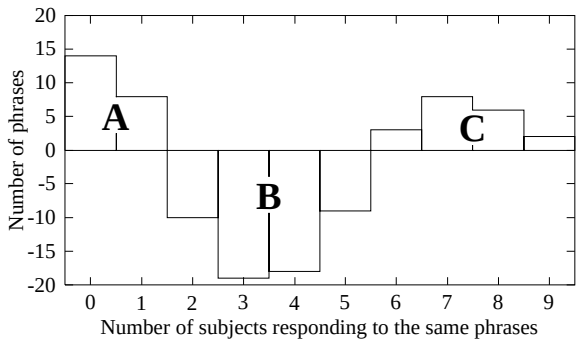

Figure 2: Distribution of pause-bounded phrases classified by the number of responding subjects (difference from binomial distribution). The y-axis represents the number of phrases falling within each class on the x-axis. The n-th class is composed of the phrases to which n subjects responded.

neutral contexts. The first two groups will be used for a machine learning method described in Section 5.

As a matter of course, these contexts chosen above may not be the same as the contexts found in the corpus because of the difference between experimental and conversational situations, of the different roles of the subjects as a third party and the participant of a dialogue, and of the accessibility to the discourse contexts. Therefore, we call those backchannels obtained by the experiment BLRs (backchannel-like responses) and temporarily distinguish them from real backchannels.

4.4. Comparison with Backchannels in the Corpus

Since BLRs are not necessarily the same as real backchannels, it would be important to examine whether or not they are quite different from backchannels in the corpus. If BLRs were completely different from real backchannels, the proposed method for detecting the context of backchannel responses would not work well for real world dialogues.

To see the difference between BLRs and real backchannels, we counted the number of backchannels found in the corpus falling within each group of BLRs. The positive group (group A) included 33 cases of real backchannels (76.7%), the neutral group (group B) 8 cases (18.6%), and the negative group (group C) 2 cases (4.7%), respectively. Thus, we can say that BLRs are not so different from real backchannels but, rather, they are quite similar. Therefore, there seems no reason to distinguish BLRs from real backchannels.

5. DETECTING THE CONTEXT OF BACKCHANNELS

In this section, we identify prosodic features which best discriminate the two groups of contexts, positive and negative contexts, of backchannels, that have been categorized in the previous section. We perform decision tree machine learning method to automatically obtain rules for discrimination, and evaluate the performance on our spoken dialogue corpus.

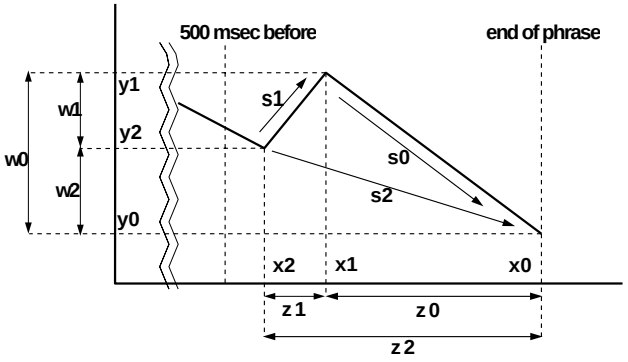

Figure 3: Prosodic features used for learning.

5.1. Prosodic Features

Several researchers noted that syntactic features as well as prosodic features can be used as cues for detecting the context for backchannels [2, 5]. We, however, do not use either lexical features (e.g., `ne' as sentence final or interjectory particles) or syntactic features (e.g., termination of grammatical clauses), since they require speech recognition and parsing. We use only prosodic features, which can be extracted without extensive labor.

We utilize prosodic features such as entire duration of pause-bounded phrases, peak levels and gradients of fundamental frequency (F0) and energy, and their temporal changes. As suggested by Maynard [5], in Japanese conversations, backchannels frequently appear during pause. Thus, we concentrate on the above prosodic features in the region of 500 msec before the end of pause-bounded phrases.

5.2. Feature Extraction

Firstly, for each pause-bounded phrase in the training data, we extracted F0 and energy by using ESPS/Waves+ software. Secondly, we approximated the final 500 msec region of the phrase by a cubic curve using least squares method. Thirdly, we obtained the coordinates of three vertices of the curve, which appear inside the 500 msec region, and other relevant parameters shown in Figure 2. In addition to the parameters in Figure 2, we also used $s1/s0\ (=t0)$, $s1-s0\ (=t1)$, and $full\_length$, which is the entire duration of the phrase. The calculation was also repeated with energy.

5.3. Obtained Rules

By using C4.5 decision tree software [7], we obtained rules for discriminating the positive and negative contexts for backchannels. Table 1 shows the rules and their coverage and accuracy. These rules are tested sequentially from the top to the bottom; the rule #0 at the bottom is the default rule, which is applied when no previous rules have been fired.

The intuitive interpretation of these rules can be summarized as follows:

- The rules for the positive contexts (#15 and #13) indicate that rise or rise-fall intonation at the end of pause-bounded phrases tend to be followed by backchannels. These rules

#	Rule	Decision	Coverage	Accuracy
2	$z2_{F0} \leq 376$ msec $w2_{F0} > -73.71$ Hz $full\_length \leq 724$ msec	negative	25.5%	98.1%
14	$y0_{F0} > 105.2$Hz $s0_{energy} \leq -40.41$ $y0_{F0} > 105.2$ Hz	negative	11.8%	91.7%
15	$t0_{F0} \leq -0.5712$ $s0_{energy} > -40.41$ $full\_length \leq 724$ msec $y0_{F0} \leq 105.2$ Hz	positive	4.9%	100%
13	$t0_{F0} > -0.4337$ $full\_length > 724$ msec	positive	26.9%	90.6%
1	$w2_{F0} > -73.71$ Hz	positive	4.9%	70.0%
0		positive	27.0%	54.5%

Table 1: Obtained rules and their coverage and accuracy. The parameters in the rules correspond to those in Figure 3. Parameters for F0 and energy are distinguished by the subscripts.

cover 31.8% of the training data, and the accuracy is over 90% when they are applied. (The rule #1 is hard to interpret.)

- The rules for the negative contexts (#2 and #14) indicate that short duration with sudden fall in loudness and flat intonation at the end of pause-bounded phrases are rarely followed by backchannels. These rules cover 37.3% of the training data, and the accuracy is over 90% when they are applied.

- The default rule is "respond with backchannels," which, however, results in very low accuracy (about 50%).

These results partly support the heuristics used in the previous studies. Our research, however, is beyond them in that the rules are derived from the data, which are expected to capture the essentials of the phenomena in more depth, and in that they show the significance of negative contexts, i.e., inhibitory cues for backchannels, which have not been discussed in detail so far.

The low accuracy of the default rule suggests the lack of sufficient negative cues among other reasons. We should try other kinds of features in the future study.

5.4. Evaluation of the Performance

We evaluated the performance of the proposed method by using cross-validation. Table 3 shows the recalls and precisions of the positive and negative cases. The recall and precision of the positive contexts are 77.6% and 69.7%, respectively; the recall and precision of the negative contexts are 68.9% and 76.8%, respectively. The overall error rate is 27.0% on the average.

Since we used pre-delimited pause-bounded phrases as data, our result is not directly comparable to those of Ward (recall: 53%, precision: 33%) and Okato et al. (recall: 77%, precision: 33%). However, the proposed method is original and very promising.

Recently, Koiso et al. [4] reported an error rate of 18.2% for the inside data using decision tree with prosodic features. Although they do not provide the results for the outside data and their research goal is quite different from ours, we believe that their results also encourage the approach outlined in this paper.

	Recall	Precision
Positive contexts	77.6%	69.7%
Negative contexts	68.9%	76.8%

Table 2: Recalls and precisions for cross-validation.

6. CONCLUSION

In this paper, we proposed a new method for detecting the context for backchannel responses by using only prosodic features. In this method, prosodic features are extracted and processed by machine learning algorithm to obtain rules for detecting the contexts of backchannels. We achieved a considerably high accuracy in the evaluation on our spoken dialogue corpus. In the future work, we wish to carry out further research with bigger corpus.

7. ACKNOWLEDGEMENTS

We would like to thank Prof. Yuji Matsumoto for providing us an opportunity to conduct this research. We also would like to thank Hanae Koiso and Yasuko Fukuda for their valuable suggestions and technical help, and Maruf Hasan for proofreading the paper.

8. REFERENCES

1. Jean Carletta, Amy Isard, Stephen Isard, Jacqueline Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. The reliablity of a dialogue structure coding scheme. *Computational Linguistics*, 23:13–31, 1997.
2. Sachiko Imaishi. The use of aizuchi in natural Japanese discourse (in Japanese). Bulletin 13, Osaka University, pp. 107–121, 1994.
3. Anne Johnstone, Umesh Berry, and Tina Nguyen. There was a long pause: influencing turn-taking behaviour in human-human and human-computer spoken dialogues. *Int. J. Human-Computer Studies*, 41:383–411, 1994.
4. Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogues. *Language and Speech*, to appear.
5. Senko Maynard. *Japanese conversation: Self contextualization through structure and interactional management*. Ablex Publishing Corporation, 1989.
6. Yohei Okato, Keiji Kato, Mikio Yamamoto, and Syuichi Itahashi. Insertion of interjectory response based on prosodic information. In *IEEE Workshop Interactive Voice Technology for Telecommunication Applications (IVTTA-96)*, pp. 85–88, 1996.
7. Ross J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1992.
8. Nigel Ward. Using prosodic clues to decide when to produce back-channel utterances. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP96)*, pp. 1728–1731, 1996.