Japanese conversation corpus for training and evaluation of backchannel prediction model

Hiroaki Noguchi*, Yasuhiro Katagiri†, Yasuharu Den‡

*Graduate School of Advanced Integration Science, Chiba University
1-33 Yayoicho, Inage-ku, Chiba 263-8522 Japan
†Department of Complex Systems, Future University Hakodate
116-2 Kamedanakano-cho,Hakodate, Hokkaido 041-8655 Japan
‡Faculty of Letters, Chiba University
1-33 Yayoicho, Inage-ku, Chiba 263-8522 Japan

*acfa3943@chiba-u.jp, †katagiri@fun.ac.jp, ‡den@L.chiba-u.ac.jp

Abstract

In this paper, we propose an experimental method for building a specialized corpus for training and evaluating backchannel prediction models of spoken dialogue. To develop a backchannel prediction model using a machine learning technique, it is necessary to discriminate between the timings of the interlocutor's speech when more listeners commonly respond with backchannels and the timings when fewer listeners do so. The proposed corpus indicates the normative timings for backchannels in each speech with millisecond accuracy. In the proposed method, we first extracted each speech comprising a single turn from recorded conversation. Second, we presented these speeches as stimuli to 89 participants and asked them to respond by key hitting whenever they thought it appropriate to respond with a backchannel. In this way, we collected 28983 responses. Third, we applied the Gaussian mixture model to the temporal distribution of the responses and estimated the center of Gaussian distribution, that is, the backchannel relevance place (BRP), in each case. Finally, we synthesized 10 pairs of stereo speech stimuli and asked 19 participants to rate each on a 7-point scale of naturalness. The results show that backchannels inserted at BRPs were significantly higher than those in the original condition.

Keywords: backchannel, spoken dialogue, machine learning

1. Introduction

In human–human conversation, a speaker does not usually make utterances in one breath; instead, he or she produces it progressively in accordance with the listener's responsive behaviors. In Japanese conversation, in particular, listeners frequently respond during the speaker's utterance with brief tokens, called "aizuchi," such as "hai" and "ee" ("uh-huh" and "yeah" in English). Aizuchi is a type of verbal backchannel feedback that conveys agreements, willingness to let the speaker continue his/her utterance, or attention to the speaker (Maynard, 1989).

It has also been reported that backchannels, when appropriately used, encourage a speaker to talk in a more fluent and lively manner (Sugito, 1993). Therefore, implementing backchannel function can make spoken dialogue systems easier to use. In the last two decades, several researchers, aiming to realize backchannel prediction models, have been trying to find prosodic and syntactic features in speakers' speeches that characterize the adequate timing for backchannel responses (Kawamori et al., 1994; Okato et al., 1998; Noguchi and Den, 1998; Ward and Tsukahara, 2000; Noguchi et al., 2001; Kitaoka et al., 2005; Nishimura et al., 2007; Kamiya et al., 2010).

The majority of these studies utilize spoken dialogue corpora and simply employ the timing of backchannels occurring in the corpora as the basis for their empirical studies. For several reasons, however, such data are not suitable for training and evaluating backchannel models. Several studies develop specialized corpora for training and evaluating of backchannel prediction model, however, they also have problems.

In this paper, we focus on the timings people respond with backchannels. We first review the characteristics of backchannels and argue the requirements of a corpus that can be used for modeling backchannels. Next, we propose a method to construct a Japanese conversation corpus that satisfies these requirements. Finally, we evaluate the constructed corpus using a naturalness-rating experiment.

2. Requirements for a backchannel corpus

2.1. Characteristics of backchannels

Normativity: Backchannels are not randomly produced. Native speakers commonly seem to know places at which they may, and should, produce backchannels. In a sense, these places are normative. Noguchi et al. (2001) named such places "backchannel relevance places" (BRPs) by an analogy to the well-known notion of "transition relevance places" proposed by Sacks et al. (1974).

Optionality: Backchannels are considered to be optional responses (Maynard, 1989), and the frequency of backchannels in our corpus varies by speaker. Table 1 shows the frequency of co-occurring backchannels from two listeners, with a time-lag threshold of 200 or 300 ms, in the Chiba three-party conversation corpus (Den and Enomoto, 2007). Although these lags are wide as mentioned below, the rate of co-occurrence is not high. Thus, listeners do not behave in the same way, even in the same context.

Recorded conversations may not contain all backchannels that might have occurred (Noguchi and Den, 1998). Simi-

Table 1: Frequency of co-occurring backchannels in a three-party conversation

Threshold	No. of co-occurring backchannels	Total no. of backchannels
200 ms	84	849
300 ms	106	

larly, the contexts that lack backchannels in recorded conversations are not always those where people should not produce backchannels. Therefore, statistical models using only contexts that are identified in a corpus as positive and negative examples would result in very low accuracy.

Extensity: Noguchi et al. (2001) showed that people are sensitive to the timing of backchannels; that is, different timing may yield different pragmatic meanings. However, they also stated that BRPs are not *points* but *places* with some duration; people consider backchannels occurring within a certain range as carrying the same meaning. Through a psychological experiment in which participants were asked to rated the naturalness of paired stimuli which varied the timing of backchannels, they found that such range had extensity longer than 250 but shorter than 400 milliseconds. Though the length of BRP may vary depending on preceding context, it shows that a BRP is a place, not a point.

2.2. Requirements

Considering the above characteristics of backchannels, there are many problems in using actual backchannels in a corpus for training and evaluating backchannel prediction models. Here, we argue two requirements for a corpus that can be used for modeling backchannles.

First, considering considerable variance across speakers, the corpus must provide the likelihood of BRPs, instead of only presenting the discrete choice between presence and absence. Furthermore, the likelihood should be provided for any moment during a speaker's utterance. To minimize presuppositions about linguistic and paralinguistic cues of BRPs, the BRPs should be described independent of any features of the speaker's speech, such as pause and intonation.

Second, the corpus must identify the BRP to which every time point in a speaker's utterance belongs. There may be two or more BRPs for a single utterance, but a backchannel occurring at a given moment in the utterance should be associated with only one BRP among them. If we can identify all the time points belonging to a BRP, we can obtain the distribution of BRPs along the time axis, which enables us to seek the context that induces Backchannels assuming that such a context is present at a time point before the center of distribution.

2.3. Related works

Maynard (1989) observed that in Japanese conversations, speakers often provide cues for inducing backchannels from listeners at or around the ends of pause-bounded phrases; she referred to such places as "backchannel contexts." Several researchers followed her theory and investigated these cues around the pre-pausal position (Maynard, 1993; Noguchi and Den, 1998; Koiso et al., 1998; Kitaoka et al., 2005). However, backchannels often overlap with a speaker's speech. It seems that certain types of linguistic and paralinguistic cues, other than pre-pausal ones, appearing in the middle of the speech may also constitute backchannel contexts. Those corpora that are annotated based on pause-bounded phrases do not meet the first requirement.

Ward and Tsukahara (2000), through examination of 80 minutes recorded natural conversations by 24 Japanese native speakers, concluded that a low pitch region continuing longer than 110 milliseconds after a speech longer than 700 milliseconds is a typical case for backchannel. They reported that a performance of their rules was a coverage of 49% and was an accuracy of 29%. They stated 44% of incorrect predictions were cases where a back-channel could naturally have appeared, and considered that such cases were caused by inter-speaker differences in back-channel behavior. Thus, their corpus does not meet the first requirement. Other corpora using recorded natural conversations (e.g. (Nishimura et al., 2007)) do not meet the first for the same reason.

Kamiya et al. (2010) constructed a backchannel utterance corpus using a experimental method; first, they presented speech stimuli to four participants and asked them to respond with backchannels as frequent as they could, and second, they compiled and manually associated participants' responses with either of a pause boundaries or a morphological segment in each stimuli. Thus, constructed on the strong presupposition about backchannel cues, it does not meet the first requirement.

Noguchi et al. (2001) constructed a specialized corpus for backchannel prediction models. They collected responses using laboratory experiments in which participants were asked to respond by key-pressing to a set of speech stimuli extracted from recorded conversations. They estimated the likelihood of BRPs at every moment of speech stimuli applying Gaussian filter with the standard deviation of 300 milliseconds, which is provisional length of a BRP. However, the distribution of the likelihood thus obtained did not always show a sharp peak and could not reliably identify a BRP. Thus, though this corpus meets the first requirement, it does not meet the second.

In the following section, we detail the experimental method of Noguchi et al. (2001) and propose a improved method for estimating BRP likelihood.

3. Proposed method

3.1. Collecting backchannel response through a laboratory experiment

Following (Noguchi et al., 2001), we collected backchannels using a laboratory experiment to create a corpus that satisfied both our requirements.

Participants: Eighty-nine college and graduate students, all native speakers of Japanese.

Material: The spoken dialogue corpus used in this paper was collected at Nara Institute of Science and Technology under the following conditions:

- face-to-face dyadic conversations
- free discussion on the topic chosen by the participants from a pre-determined list
- recording done in a soundproof room
- headset type microphones used (without headphones)
- recording done on separate channels and sampled in high quality at a rate of 20kHz

Forty minutes of dialogues in total by three pairs of participants were transcribed. Speech materials were divided into pause-bounded phrases delimited by pauses longer than 100 ms, yielding 1875 such phrases.
From this corpus, we selected, 176 stimuli, each consisting of several pause-bounded phrases and constituting a single dialogue act. We excluded cases that were deemed to difficult to understand or listen to, that were too short to elicit an adequate response, or that contained only one pause-bounded phrase. Then the average number of pause-bounded phrases contained in a stimulus was 2.91, and the average length of a stimulus was 4.8 sec.

Procedure: Participant were asked to respond to stimuli by pressing the space bar whenever they thought it appropriate to respond with a backchannel. Each participant was presented with all stimuli in a random order and without discourse contexts.

Results: The number of responses obtained from 89 participants for 176 stimuli was 28983. For each stimulus, we compiled the participants' responses on the time axis. Figure 1 shows an example of the distribution of responses; the horizontal axis indicates the elapsed time from the beginning of the stimulus, and the vertical lines indicate the timing of each responses. It was found that there were certain dense zones in which several participants commonly responded within a small time range. On the other hand, in the remainder of the stimulus participants did not commonly respond. Thus, the likelihood of BRP should be estimated as high in the former but low in the latter.

3.2. Estimating likelihood of BRP

It is generally accepted that a listener responds with a backchannel upon recognizing a backchannel context. Since the response latency differs by person, it seems reasonable to consider that the distribution of responses to a single context conforms to a Gaussian distribution. Moreover, all the speech stimuli used in the present experiment were considerably long (mean duration = 4.8 sec) and may contain more than one backchannel context. Thus, it seems reasonable to assume that the distribution of responses to a single stimulus conforms to a mixture of Gaussian distributions.

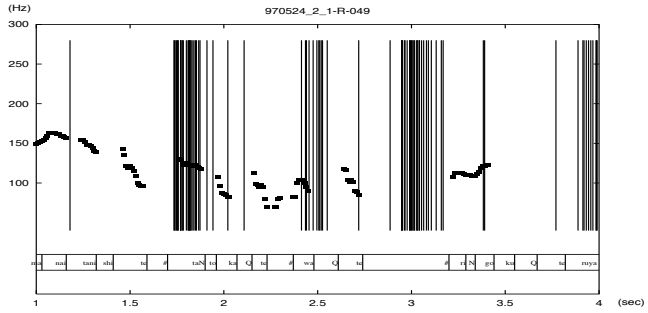

Figure 1: Example of the distribution of responses. The F_0 contour of the speech stimulus is superimposed.

To estimate the likelihood of BRP, we used "mclust," a package of the R language (Team, 2013) designed for model-based clustering, classification, and density estimation based on Gaussian mixture modeling (Fraley et al., 2012). "mclust" estimates the number of clusters, the component Gaussian models, and the probabilistic density at every time point. Here, we consider the estimated Gaussian models as BRPs and the probabilistic density as their likelihoods.
Figure 2 shows the result for the same speech stimulus as in Figure 1. The responses collected during the experiment are clustered into three groups, demonstrating that there are three backchannel contexts in the stimulus. The figure also shows the density at every moment in the stimulus, which constitutes the likelihood of BRPs.

4. Evaluation by naturalness-rating experiment

To evaluate the adequacy of the obtained BRPs, we conducted an additional experiment in which participants were asked to compare pairs of speech stimuli to assess which paired stimuli is more natural. One of the pair backchannels was produced according to the obtained BRP likelihood, while the other was an actual backchannels from the corpus.

Participants: Nineteen college students, all native speakers of Japanese.

Stimuli: Ten paired stimuli were presented to each participant. Each stimulus was a speech segment randomly selected from the corpus, and it received responses with synthesized backchannels with varied frequencies and timings. In one of a paired stimuli, backchannels were produced according to the likelihood of BRPs obtained in subsection 3.2. (the BRP condition). In the other of a paired stimuli, the synthesized backchannels were produced at the timing of the actual backchannels in the original corpus (the original condition). To eliminate the variable influence different forms of backchannels may have, the same form of backchannel, "um," was always used.

Procedure: Participants rated the naturalness of each paired stimuli on a 7-point scale, evaluating it on the basis of three aspects: (i) the intensiveness of the listener, (ii) ease with which the speaker talk, and (iii) the liveliness of the interaction. In addition, the participants were able to repeatedly play and compare paired stimuli.

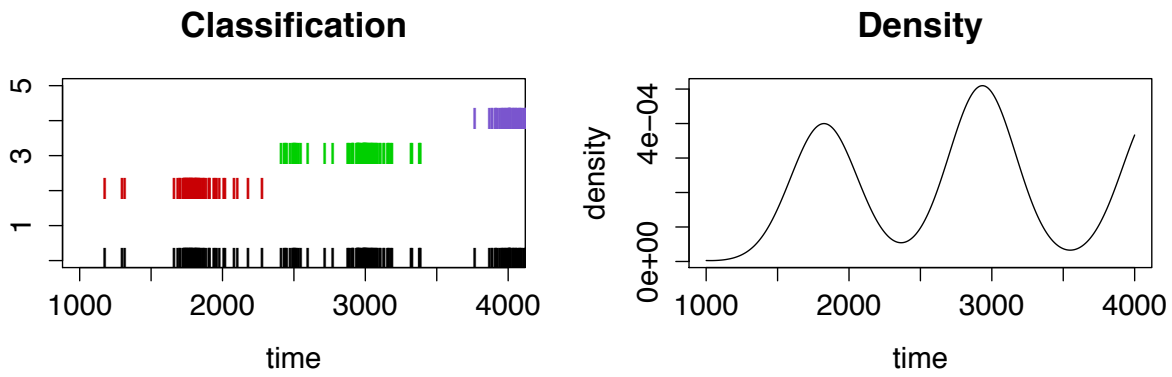

Figure 2: Estimated clusters and densities for responses to the sample stimulus.

Evaluation: We used Nakaya's variant of Scheffe's method of paired comparison (Nakaya, 1970) to statistically test which of the two conditions was rated as more natural. Scheffe's method estimates the ranking among these sets through pairwise comparison, and Nakaya's variation enables participants to repeatedly and alternatively compare pairs as they want.

Result: The BRP condition was rated significantly higher than the original condition in all three aspects of naturalness. One explanation for this is that in the original corpus, the listeners mutually keep eye contact and therefore do not need to frequently respond with backchannels. In fact, the stimuli in the original condition contain fewer backchannels than those in BRP.

5. Discussion

In this paper, we discussed the requirements for a corpus that can be used for training and evaluating backchannel prediction models as well as proposed a new method for constructing such a corpus.

We used only 176 single-turn speech stimuli; such a limited corpus may not be sufficient to cover the variation of linguistic and paralinguistic backchannel cues. The proposed method of collecting backchannel responses, however, is simple and it is easy to expand the corpus by conducting additional experiments with more speech stimuli. Similarly, it is easy to enhance the reliability of the BRP likelihood estimation by increasing the number of participants, especially since there are no restrictions regarding the nature of participants.

Finally, owing to the simplicity of the experiment, contrastive studies in different languages are also possible.

6. References

Yasuharu Den and Mika Enomoto. 2007. A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation. In T. Nishida, editor, *Conversational informatics: An engineering approach*, pages 307–330. John Wiley & Sons.

Chris Fraley, Adrian E. Raftery, T. Brendan Murphy, and Luca Scrucca. 2012. MCLUST: Software for model-based cluster analysis mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report 597, University of Washington, Department of Statistics.

Yuki Kamiya, Tomohiro Ohno, Shigeki Matsubara, and Hideki Kashioka. 2010. Construction of back-channel utterance corpus for responsive spoken dialogue system development. In *Proceedings of the seventh International Conference on Language Resources and Evaluation*, pages 2414–2149.

Masahito Kawamori, Akira Shimazu, and Kiyoshi Kogure. 1994. Roles of interjectory utterances in spoken discourse. In *Proceedings of the third International Conference on Speech Communication and Technology*.

Norihide Kitaoka, Masashi Takeuchi, Ryota Nishimura, and Seiichi Nakagawa. 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):220–228.

Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and speech*, 41(3-4):295–321.

Senko Maynard. 1989. *Japanese conversation: Self contextualization through structure and interactional management*. Ablex Publishing Corporation.

Senko K. Maynard. 1993. *Convesational Analysis (in Japanese)*. Kuroshio, Tokyo.

S Nakaya. 1970. Variation of schffe's paired comparison. In *Proceedings of 11th Sensory Evaluation Convention (in Japanese)*, pages 1–12.

Ryota Nishimura, Norihide Kitaoka, and Seiichi Nakagawa. 2007. A spoken dialog system for chat-like conversations considering response timing. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue*, pages 599—–606.

Hiroaki Noguchi and Yasuharu Den. 1998. Prosody-based detection of the context of backchannel responses. In *Proceedings of the fifth International Conference on Spoken Language Processing*, pages 487–490, Sydney.

Hiroaki Noguchi, Yasuhiro Katagiri, and Yasuharu Den. 2001. A proposal of building a BRP-likelihood corpus (in Japanese). *Technical Report of the Japanese Society for Artificial Intelligence*, SIG-SLUD-A101:25–32.

Yohei Okato, Keiji Kato, Mikio Yamamoto, and Shuichi Itahashi. 1998. System-user interaction and response strategy in spoken dialogue system. In *Proceedings of*

the fifth International Conference on Spoken Language Processing.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

Miyoko Sugito. 1993. Effective timing and character of involved conversation and backchannels. *Journal of Japanese Linguistics*, 12(4):11–20.

R Core Team, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 32(8):1177–1207.