AN EXPERIMENTAL VERIFICATION OF THE PROSODIC/LEXICAL EFFECTS ON THE OCCURRENCE OF BACKCHANNELS

Hiroaki Noguchi\(^1\) \(^\dagger\), Yasuhiro Katagiri\(^1\) and Yasuharu Den\(^1\)

\(^1\)ATR Media Integration and Communications Research Laboratories
2-2 Hikaridai, Seikacho, Kyoto, 619-0288 Japan
\(^\dagger\)Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayamachō, Ikoma, Nara, 630-0101 Japan
e-mail: \{noguchi, katagiri\}@mic.atr.co.jp, den@is.aist-nara.ac.jp

ABSTRACT

Japanese backchannel utterances have recently been studied to incorporate their conversational and social functions into spoken dialogue interfaces. Most of the studies employ corpus-based methodologies to elucidate conditions under which backchannel utterances occur. We propose, in this paper, an experimental method to identify backchannel-contexts through controlled manipulation of local prosody of conversational utterances by incorporating speech synthesis technologies. We found that backchannel-facilitating contexts have to be characterized both in terms of the lexical features and the local prosody, whereas backchannel-suppressing contexts can be partially identified in terms of the phrase-end prosody.

1. INTRODUCTION

For several reasons, a speaker usually does not utter his/her whole utterance at one time, but instead, completes the utterance progressively by inserting several phrasal pauses. In Japanese conversation, participants often respond during utterances with backchannel feedback, which is a social interaction that conveys slight agreement, a willingness to let the speaker to continue his utterance, or the idea that the listener is paying attention to the speaker\cite{6}. It has also been reported that backchannels, when appropriately used, encourage a speaker to talk more fluent and lively manner\cite{9}; therefore it is generally accepted that addition of a backchannel function can make a spoken dialogue system easier to talk with. Several researchers have tried to find ways of detecting adequate timings for backchannels by analysis based on spoken dialogue corpora and some prosodic and lexical clues have been found.

Although the corpus-based methodology seems appropriate, it does not provide sufficient evidence that prosodic features are not capable of detecting adequate timing for backchannels while it does say that utterances with prosodic features are likely to be followed by backchannels in corpora, the occurrence of these features may nonetheless only constitutes a pseudo-correlation with the occurrence of backchannels; that is, prosodic features simply accompany the actual clues a speaker uses for backchannel timing. It could also be that only a simple feature like prosody is able to explain the occurrence of backchannels whereas corpus analysis could not separate it from other accompanying features. In particular, from the engineering viewpoint, the latter case is very beneficial because prosodic features are relatively easy to handle and may enable the development of automatic backchannel response systems. Therefore, it is important to confirm whether prosodic features can actually be used for backchannel timing. A corpus-based analysis is not appropriate for this purpose, and, accordingly, we need an alternative methodology.

In this paper, we propose an experimental methodology that verifies the actual clueness of prosody itself as a marker of adequate context for backchannels by using speech synthesis technology. We also show sufficient evidence that a listener cannot determine an adequate context for backchannels by listening only to the prosody. Section 2 reviews related works on clues for detecting the context of backchannels in Japanese conversation. Section 3 describes our hypothesis on backchannels, on which the experiment is based. Section 4 describes the procedure, environment, subjects and results of the experiment. Section 5 discusses our results, and section 6 is a brief conclusion.

2. BACKCHANNEL CONTEXT

Maynard observed that in Japanese conversations, a speaker often provides cues for inducing backchannels from the listener at or around the ends of pause-bounded phrases, and called such timing as “backchannel context”\cite{6}. She also suggested that sentence final and interjectory particles serve as lexical clues, e.g., “ne” in Japanese, when followed by a pause. Since then, many researchers have speculated about the factors that determine backchannel contexts and have proposed various prosodic and lexical features that could be used as cues for backchannels; a pause-bounded phrase which has both either rise-fall or fall intonation and a interjectory particle at its end\cite{9, 3, 7}, one with fall intonation\cite{9, 7}, one with rise-fall intonation\cite{7}, one with both rise-fall intonation and a case particle\cite{5}, and one having certain pitch pattern\cite{8}. In addition, two cues for when a listener should not respond have been suggested; one with flat intonation at its end\cite{9, 3, 7} and one with a case particle. Ward found a heuristics that states that a listener should respond with a backchannel when a low pitch region in a speaker’s speech lasts longer than 150 msec. He concentrated on global prosody\cite{10}, whereas the other reports focused on the local view. Figure 1 shows how these reports are related.

Most of the research has focused on the local clues and can be classified into three categories: one that uses only lexical features at the end of an utterance as clues\cite{6}, those that use only prosody\cite{9, 8, 7}, and those that use combinations of prosody and lexical features at the end of an utterance\cite{5, 3}. In this paper, for the purpose of comparison with the previous methods, we utilize features...
3. HYPOTHESIS

Roughly speaking, the central idea of previous works on backchannels has been that a participant speaks an utterance while giving clues like lexical features at the end in order to indicate the timing for backchannels to the listener and that the listener responds with backchannels by sensing those clues. Since this paper examines a single clueness of local prosody for backchannel context, we have to show that 1) a difference in prosodic features make a difference in the occurrence of the following backchannels and that 2) prosodic features account for the occurrence of observed backchannels more precisely than do any other type of features, such as lexical categories. To verify that both of the above conditions are met at the same time, we conducted a psychological experiment in which a subject is asked to respond by hitting a button whenever he/she thinks it appropriate to ‘backchannel’ while listening to a utterance stimulus. The pitch contour is re-organized into one of three types of intonation contours at the utterance’s end. The utterance has one of 3 kinds of parts of speech in its end, and then we count which pause-bounded phrase within a stimulus are followed by a backchannel. This experiment allows us to obtain the frequency of occurrence of backchannels in each context. Fortunately, a speech synthesis technique called “STRAIGHT”[4] enables us to manipulate the prosodic feature relatively easier than to manipulate the lexical or contextual conditions.

4. EXPERIMENT

We conducted a psychological experiment in which subjects were asked to respond with backchannels to utterance stimuli whenever they felt doing so was appropriate. By using a speech synthesis technique, we could directly evaluate prosodic and lexical features on backchannels by observing variations in the occurrence of backchannels in a variety of prosodic and lexical conditions.

4.1. Method

Subjects 15 college students, all native speakers of Japanese.

Materials We used a spoken dialogue corpus that was collected at the Nara Institute of Science and Technology (hereafter, NAIST) under the following conditions:

- face-to-face dyadic conversations
- fully spontaneous conversation
- 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 20 kHz

Speech materials were divided into pause-bounded phrases delimited by pauses longer than 100 msec. Figure 2 shows an excerpt from the corpus with an example of backchannels.

Stimuli We produced utterance stimuli for the experiment by the following procedures. First, we selected, from the corpus, a set of utterances, each of which consisted of several pause-bounded phrases and constituted a single conversational move[1] which constitutes an exchange. These sets could be easily understood. Such utterances that were too short to respond to or that contained only one pause-bounded phrase were not included. Note that our definition of backchannels follows convention; only responses to move-internal pause-bounded phrases are considered to be backchannels. Second, we classified utterances into three types by phrase-end lexical categories such as ‘interjectory particle,’ ‘case particle,’ and ‘adverb,’ which have repeatedly been reported as being cues for backchannels in previous works, and obtained 12 utterances for each type of lexical categories in the phrase-end. Third, we used STRAIGHT[4], a vocoder that produces natural and clear speech waveform, to re-synthesize each utterance into three variations of speech utterances that differed in pitch contour in the final mora, ‘rise-fall,’ ‘fall,’ or ‘flat.’ In total, we obtained 108 speech stimuli.

The F0 values of the re-synthesized speech were calculated as follows:

\[
\begin{align*}
\text{rise-fall intonation} & : F_0_i = F_0_0 + 50 \times \sin \frac{2\pi t}{T} \\
\text{fall intonation} & : F_0_i = F_0_0 - 20 \times \sin \frac{2\pi t}{T} \\
\text{flat intonation} & : F_0_i = F_0_0
\end{align*}
\]

\(F_0_0\) was the value at the beginning of final mora. Furthermore, we make each pause length equal at 300 msec.

Procedure We conducted the experiment as follows:

step 1 We presented the 36 of the 108 utterance stimuli to each subject in random order without discourse contexts.

step 2 Each subject responded to utterances by hitting a keyboard whenever he/she felt it was adequate timing for a backchannel.

step 3 We counted the number of subjects responding at each observation point, that is, each phrase pause. Each pause-bounded phrase within a stimulus had to be judged as to whether it was to be followed by a backchannel by a subject within 100 to 400 msec after the end of that phrase; otherwise, we did not count the subject’s response as...
Figure 2: Excerpt from the corpus with English translation on the right column. ‘L’ and ‘R’ identify speakers. Each line corresponds to a conversational move[1] and a ‘/’ indicates a boundary between phrases. Backchannels are in boldface.

Table 2: The results of multiple comparison tests.

<table>
<thead>
<tr>
<th>Interaction</th>
<th>Statistics</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adv/P-Case &amp; P-Intj</td>
<td>F(1, 228) = 20.191</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Flat &amp; P-Intj</td>
<td>F(1, 228) = 8.267</td>
<td>p &lt; 0.01</td>
</tr>
<tr>
<td>Flat &amp; Fall/Rise-Fall</td>
<td>F(1, 228) = 5.492</td>
<td>p &lt; 0.01</td>
</tr>
<tr>
<td>Flat &amp; P-Intj &amp; P-Case</td>
<td>F(1, 228) = 11.818</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Flat &amp; P-Intj &amp; Rise-Fall</td>
<td>F(1, 228) = 5.510</td>
<td>p &lt; 0.01</td>
</tr>
<tr>
<td>Adv/P-Case &amp; Fall/Rise-Fall</td>
<td>F(1, 228) = 9.388</td>
<td>p &lt; 0.01</td>
</tr>
<tr>
<td>Adv/P-Case &amp; Rise-Fall</td>
<td>F(1, 228) = 7.403</td>
<td>p &lt; 0.01</td>
</tr>
<tr>
<td>Adv/P-Case &amp; P-Intj &amp; P-Case</td>
<td>F(1, 228) = 2.5285</td>
<td>p = n.s.</td>
</tr>
</tbody>
</table>

In the table, “Adv”, “P-Case”, “P-Intj” are abbreviations for adverb, case particle, and interjectory particle, respectively, and “<” indicates that there is a significant difference in the average number of responding subjects. Lines 1 and 2 show the trinomial relation, and so do lines 3 and 4. They mean that either prosody or part of speech has a significant effect on backchannel occurrence. Furthermore, lines 7 to 7 show that an utterance having both a P-Intj and Fall/Rise-fall intonation at its ends is significantly followed by more backchannels than are the other conditions.

These results show that a pause-bounded phrase with both an interjectory particle and either a rise-fall or fall intonation at its end is more likely to be followed by a backchannel and that a pause bounded phrase with either an adverb or continuous F0 at its end is less likely to be followed by a backchannel, regardless of other features. This result coincides with those of previous research [9, 3]. From it, we can deduce that the timing that people think is not appropriate for response can be partly predicted with only prosody. However, it is not possible to predict the appropriate timing for backchannels with only prosodic clues at the end of phrases.

5. DISCUSSION

Although the results of the experiment in the previous section elucidate the effects of prosody and part-of-speech on the occurrence of backchannels, there might be other variables that explain it more precisely. Here, we focus on other possible factors. First, it could happen that the subjects could not respond to pause-bounded phrases that are too short. Therefore, we calculated the Pearson’s product-moment correlation coefficient between the length of an utterance stimulus and the frequency of backchannels. However, there was no significant correlation ($F(1, 34) = 2.5285, p = n.s.$) Second, it could happen that when the observed point is at the very beginning of a stimulus, subjects cannot respond, whereas with an observed point more distant from the be-
As we have pointed out; corpus based analysis, which correlates a certain kind of phenomena with certain characteristics that precede them, shows only that the phenomena are likely to follow those utterances which have a certain cues like prosody. Our alternative experimental methodology, however, revealed that one can not detect adequate timings for backchannels by using only prosody or part of speech in the utterance, and that, instead, both types of the cues are necessary.

These results are in partial agreement with some previous works by Imaishi [3] and by Sugito [9], in which an utterance which has both rise-fall intonation and interjectory particle at its end is likely to be followed by a backchannel, and an utterance which has flat intonation at its end is less likely to be followed by a backchannel. Since, almost of all of the result obtained in this study have been suggested in previous research. One may think that this study has nothing to contribute; definitely, not true. For instance, while it is known that rise-fall intonation is likely to have occurred when a listener responds with a backchannel, it is not clear that features that are likely to accompany with rise-fall intonation have actual clueness for backchannels or not. In fact, since detection with only prosodic features can achieve to a certain level of accuracy [8, 7] as shown in the Table 1 that prosody by itself has significant effect on backchannels, non-prosodic features have largely been ignored, even among researchers developing spoken dialogue system applications. However, our experiment clearly shows that both prosodic and lexical features are required to detect the timing.

In this study, we focused on the local prosodic and lexical features and found that prosody by itself at an utterance’s end cannot aid detection. However, this does not mean that we are unable to detect timing of backchannel by prosody itself. We think lexical features are one of the contextual features, which constitute more broader structures and whose arrangement can reflect the global structure. Furthermore, prosody can also reflect the global structure. Therefore, it may be possible to exchange the role of local lexical features with that of global prosody. We have already started to investigate this possibility.

6. REFERENCES


