Toward Adaptive Generation of Backchannels for Attentive Listening Agents

Tatsuya Kawahara, Miki Uesato, Koichiro Yoshino, Katsuya Takanashi

Abstract Backchannels play an important role in spoken dialogue, especially in attentive listening such as counseling. Appropriately coordinated backchannels help establish rapport in that kind of dialogue. We investigate whether and how synchrony is expressed by the prosodic features of backchannels with respect to the preceding speaker's utterances. By analyzing counseling dialogue, we find out correlation patterns according to the type of backchannels and prosodic features; a larger correlation is observed for reactive tokens than acknowledging tokens and for the power features than the pitch features. Based on the observations, we also conduct prediction of prosodic parameters of backchannels in order to replace the conventional conversational systems that generate monotonous backchannels.

1 Introduction

Feedback behaviors play an important role in smooth communication [1]. In speech communication or spoken dialogue, verbal backchannels, such as "okay" and "right" in English, convey feedback. Without the feedback, the speaker would be anxious whether the communication is well maintained and would feel as if talking to a "machine".

Backchannels are used to express the listener's feedback to what is uttered while suggesting that the current speaker can keep the dialogue turn. Specifically, backchannels express that the listener is listening, understanding, and agreeing to the speaker. Backchannels can also be used to express the listener's assessment such as surprise, interest and sympathy. The variety of these roles is correlated with lexical and prosodic patterns of the backchannels [2, 3].

In addition to the effect of individual backchannels, backchannels make a "rhythm" of the dialogue as a whole. By making "synchrony", dialogue partners feel comfortable in keeping the dialogue. The phenomenon is regarded as one aspect of

All authors are with School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan.

entrainment [4]. In counseling, it is crucial for a counselor to keep the client talking on his/her matter by establishing rapport. To that end, counselors make effective use of backchannels to express empathy and make synchrony in the dialogue [5].

The work presented in this paper focuses on the synchrony effect of backchannels rather than their individual role and meaning. Specifically, we investigate whether prosodic synchrony is observed in generating backchannels in counseling dialogue, and whether this information can be used to predict prosodic patterns of backchannels.

This finding would be useful for designing a new kind of spoken dialogue systems or conversational agents, which conduct attentive listening. A majority of current systems are designed for task-oriented dialogue or simple information retrieval, which assumes users have something to ask. On the other hand, conversational systems would be useful by attentively listening to particular user populations such as elderly and ill persons. In order to make smooth communication by establishing rapport, appropriate backchannel generation is critical. Note that speech recognition and understanding may not be necessary to realize this function.

In the remainder of the paper, we first review the work on analysis and generation of backchannels in Section 2 and the counseling corpus collected for this work in Section 3. Based on the corpus, we present an analysis on prosodic synchrony by backchannels in Section 4 and an experiment of predicting prosodic features of backchannels in Section 5.

2 Analysis and Generation of Backchannels

A verbal backchannel is a short response generated by the listener during the dialogue, usually at the end of utterances, without taking a turn; instead backchannels suggest that the listener does not take a turn. By this definition, backchannels are distinguished from acknowledgement and fillers, which are used to take or keep a turn in the dialogue.

In generating or analyzing backchannels, we need to determine or identify following three factors: lexical entry, timing and prosody.

2.1 Lexical entry

The lexicon of backchannels is language-dependent, and is not focus of this work.

In general, segmental patterns of backchannels are classified into two categories. One is lexical and usually same entries as acknowledging tokens such as “okay” and “right” in English and “*hai*” and “*un”* in Japanese. The other is non-lexical reactive tokens such as “uh-huh” in English and “*hu:n*” in Japanese. The acknowledging tokens are more frequently used and they indicate that the listener is listening, understanding, and agreeing to the speaker, while the reactive tokens are specially used to indicate the listener's strong reaction and assessment to what is uttered.

2.2 Timing

The timing of backchannels is usually constrained at the end of the current speaker's utterances, but whether to make a backchannel is determined by a number of factors.

There are a number of previous studies that investigated the cues of backchannels, or when to make a backchannel. As early work, Ward et al. [6, 7] pointed out the low pitch as a major prosodic cue of backchannels. Koiso et al. [8] and Noguchi et al. [9] introduced a decision tree to derive rules from prosodic and morphological patterns. Recent studies mainly focus on refinement of prosodic cues [10, 11, 12].

There are also several studies which actually implemented a dialogue system to generate backchannels using a decision tree [13, 14]. Recently, more elaborate discriminative modeling and an efficient learning mechanism using the wisdom of crowds have been introduced [15, 16].

Although timing is an important issue to generate backchannels, it is not focus of this work, either.

2.3 Prosody

The prosody of backchannels is important especially for expressing assessment with reactive tokens, and we have identified particular patterns to express interest and surprise in conversations [3, 17]. On the other hand, the general prosodic patterns of backchannels have not been carefully investigated, compared to the prosodic cues of backchannels. Actually, almost all systems that generate backchannels mentioned above use the same recorded or synthesized backchannel pattern "*hai*" or "okay", which gives a monotonic impression.

Recent work by Heldner et al. [18] showed that there is pitch similarity in the vicinity of backchannels, that is pitch of backchannels is more similar to that of the preceding utterances, compared with normal turn-taking. It suggests that pitch of backchannels is controlled to synchronize with the dialogue partner. Our work presented in this paper is to further develop this standpoint. We deal with not only pitch but also power as the stress is used mainly for para-linguistic information in Japanese. Moreover, we conduct a correlation analysis for exploring effective prediction of the prosodic features.

3 Corpus of Counseling Dialogue

In order to conduct an analysis of attentive listening and develop a prototype system of such function, we have recorded sessions of counseling dialogue. These are not real counseling, in that the subjects were asked to come to the session for dialogue data collection, not for counseling. But they were asked to talk about their real personal troubles, for example, human relationship and the career path, to a counselor. The subjects are six college students of 20 to 25 years old. We had two counselors (one male of 7-year counseling experience and one female of 4-year experience), and each took part in three sessions. All participants are Japanese native. Among six dialogue sessions, four were conducted in the face-to-face mode and two were not; the participants were in the same room but there was a screen between them.

The dialogue started with some chatting and the following counseling session lasted around 20-30 minutes. The speech was captured by a head-set microphone worn by each participant and transcribed according to the guideline of the Corpus of Spontaneous Japanese (CSJ) [19]. Backchannels were annotated separately.

The statistics of backchannel occurrence are shown in Table 1. Here we focus on backchannels made by the counselors. Many backchannels are repeated one such as "*hun hun*", and they are counted as single occurrence as long as uninterrupted. It is observed that the counselors make backchannels every 5 to 7 seconds, meaning at almost every end of the speaker's utterances. Actually, they were trained to make backchannels.

The statistics by the lexical entries show that a large majority of backchannels are acknowledging tokens such as "*hun*", "*u:n*", "*un*", "*hu:n*", "*hun hun*", "*un un un*", "*un un*" in the descending order of the occurrence count. Since it is apparently difficult to distinguish "*un*" from "*hun*" and also "*u:n*" from "*hu:n*", we deal with them collectively. They are clustered based on whether they are prolonged and the number of the repetitions, and represented, for example, by "*un* x2" or "(*un*)+".

Table 1 Statistics of backchannel occurrence in counseling dialogue

gender of counselor	face-to -face?	duration of session (min.)	occurrence of backchannels	occurrence per minute
female	yes	26.02	275	10.56
female	yes	21.15	181	8.51
male	yes	17.27	212	12.15
male	yes	22.30	293	13.01
female	no	31.20	294	9.38
male	no	21.09	332	9.44
average		23.17	265	11.42

4 Analysis on Prosodic Synchrony by Backchannels

We investigate synchrony in prosodic patterns expressed by the listener's backchannels with respect to the preceding speaker's utterances.

4.1 Prosodic Features

We focus on the prosodic features of the speaker's utterances preceding the backchannels of the counselor. There are many overlapping cases between them, but each segment of 500 msec from the end of the utterance was analyzed using the speech data captured by the head-set microphone.

Fundamental frequency (F0) was computed with a frame shift of 10 msec using wavesurfer 1.8, [1] then it was converted to the logarithm scale and normalized with the mean and the standard deviation computed per person for the entire session. The final value is referred to as z-score. Power (in dB) was also computed using wavesurfer 1.8 and normalized in the same manner.

4.2 General Synchrony in Prosodic Features

We first investigate the general tendency of synchrony, which was reported by Heldner et al. [18]. The synchrony is measured by comparing with normal turn-taking. To this end, we also computed the prosodic features (mean log F0 and power) for the beginning segment of 500 msec when the counselor takes a turn to ask some questions or make a comment.

We used 952 samples of backchannels and 279 samples of normal turn switches in this experiment. Since counselors are mostly engaged in attentive listening, their turn-taking is not frequent. We measured the distance between the prosodic feature of these segments and that of the preceding utterances (i.e. backchannel vs. preceding utterance & turn-taking vs. preceding utterance), as shown in Figure 1.

Heldner et al. [18] reported that the distance of F0 between backchannels and their preceding utterances is significantly smaller than that of normal turn switches, that is pitch of backchannels is closer to that of the preceding utterances compared with normal turn-taking. However, this phenomenon is not confirmed in our corpus. There is not a significant difference in F0 between these two cases (left graph of Figure 1). Instead, we observe a significant difference in the power feature; power of backchannels is much closer to that of the preceding utterances compared with normal turn-taking (right graph of Figure 1).

[1] `http://www.speech.kth.se/wavesurfer`

Although we do not have a clear explanation for the results, the difference in general prosodic patterns between English and Japanese and also the difference in segmental and prosodic patterns of backchannels may be attributed.

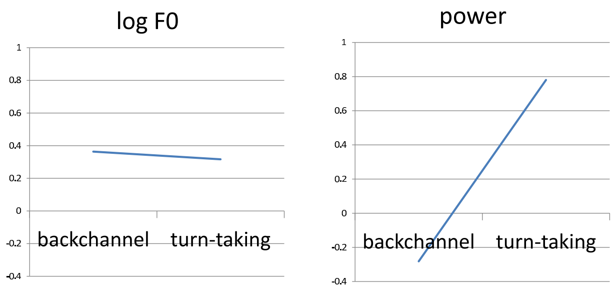

Fig. 1 Difference in prosodic features from preceding utterances (comparison of backchannels and normal turn-taking)

4.3 *Correlation of Prosodic Features between Backchannels and Preceding Utterances*

Next, we investigate the correlation of the prosodic features of backchannels and those of the preceding utterances. This will reveal more precisely whether and how synchrony is realized in generating backchannels. The analysis is conducted for each category (acknowledging tokens and reactive tokens) and for each clustered segmental pattern, but those with fewer occurrence counts (less than 25) are not used.

We compute F0 and power (Pow) in the same manner as in the previous subsection, and parameterize them with their mean, maximum (max) and range within the segment. Then, a correlation coefficient is computed between the parameter of backchannels (by listener L) and that of the preceding utterances (by speaker S). Here, we also investigate the correlation between different features, for example, power of the speaker's utterance (S:Pow) and F0 of the listener's backchannels (L:F0).

The results of the correlation analysis are presented in Table 2. Here we list those with significant correlation for the acknowledging tokens (upper part of Table 2) and those larger than 0.20 for the reactive tokens (lower part of Table 2) since the latter does not have a large number of samples. We can see more correlation patterns with regard to the power feature of backchannels (L:Pow). This confirms the result of the previous sub-section: synchrony is observed for power rather than pitch. Larger correlation patterns are observed in the power feature of repetition patterns such as "*un un*", while a small correlation is found in the F0 feature of short backchannels of "*un*" and "*u:n*". It suggests that the listener can easily control the power parameter in long backchannels.

Much larger correlation patterns are observed for the reactive tokens of "*a:*" and "*ha:*" although there are a small number of these samples. It is natural that they are used to express strong reaction to the speaker.

There are some cross-feature correlations, for example, it is suggested that power of reactive tokens is also controlled depending on F0 of the speaker's utterances.

In summary, power is adjusted to make synchrony in repeated tokens and reactive tokens, while pitch plays some role in short tokens.

Table 2 Correlation of prosodic features between backchannels and preceding utterances

segmental pattern	count	S:F0 vs L:F0	S:Pow vs L:Pow	S:F0 vs L:Pow	S:Pow vs L:F0
u:n	225	max (0.14)	max (0.14) mean (0.18)	range (0.14)	
un	361	max (0.22) mean (0.12)	max (0.23)		max (0.22) range (0.15)
un x2 *un un*	146	mean (0.20)	max (0.34) mean (0.35)		
un x3 *un un un*	169		max (0.32) mean (0.30)		
un x4 *un un un un*	117		max (0.24) mean (0.22)	max (0.19) mean (0.33)	
a:	28	mean (0.22)	mean (0.25)	range (0.38)	
ha:	27	mean (0.23)	max (0.47) mean (0.29)	range (0.28)	

Significant correlation patterns are shown with their correlation coefficients.

5 Prediction of Prosodic Features of Backchannels

Finally, we conduct prediction of the prosodic features of backchannels based on those of the preceding utterances by the speaker.

5.1 Formula of Prediction

We formulate a simple prediction model based on the correlation presented in the previous section. The model is designed to change the prosodic patterns of backchannels depending on the current speaker's utterances, as opposed to the conventional conversational systems that generate the same backchannels through the entire dialogue session. Thus, the baseline is to use the mean value of the prosodic features. The proposed model modifies it by the following formula:

$$\hat{BC_i} = \alpha \times \{ \frac{S_i - E(S)}{\sigma(S)} \times \sigma(BC) \} + E(BC) \quad (1)$$

where i is the index for the pair of the current speaker's utterance S and the listener's backchannel BC. S_i and BC_i represents a prosodic feature of the i-th utterance and backchannel, respectively. $E(S)$ and $E(BC)$ represents a mean and $\sigma(S)$ and $\sigma(BC)$ represents a standard deviation of the corresponding prosodic feature. α is a coefficient weight, and we use the correlation value defined in the previous section. We also tried to tune this weight to minimize the mean square error defined in the next sub-section, but the result was not changed so much.

The prediction model is prepared, namely all model parameters mentioned above are estimated separately for the two categories of acknowledging tokens and reactive tokens, which correspond to the upper part and the lower part of Table 2. For each category, all segmental patterns are merged to make a single model.

5.2 Result of Prediction

Prediction is performed for a prosodic parameter (mean/max of log F0 or power) of the backchannels in the counseling dialogue corpus using the same parameter of the preceding speaker's utterances, e.g. the mean power of a backchannel is predicted from the mean power of the preceding utterance.

Prediction performance is measured by the mean square error (MSE) defined below:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{BC_i} - BC_i)^2 \quad (2)$$

$$(i = 1, \ldots, N)$$

It measures the difference between the predicted value $\hat{BC_i}$ and the actual value BC_i, and N is the number of the samples. Here, the error of F0 is computed in the linear scale instead of the logarithm scale.

The square root of MSE is listed for the prosodic features of acknowledging tokens such as "*u:n*" and "*un un*" in Table 3 and for those of reactive tokens such as "*a:*" and "*ha:*" in Table 4. The total number of the samples is 586 and 64, respectively. Note that the baseline model simply uses the mean value by setting α=0 in Equation (1).

A larger improvement is confirmed in the prosodic features of reactive tokens, since a larger correlation is observed for them as in Table 2. Moreover, the range (standard deviation) of the prosodic features of reactive tokens is generally larger than that of acknowledging tokens [3]. Therefore, it is easier to control or synchronize the prosody in reactive tokens, depending on the preceding utterances.

Table 3 Prediction result of prosodic features of acknowledging tokens {*u:n*, (*un*)+} (square root of MSE)

	baseline	prediction
F0 max (Hz)	30.1	29.4
F0 mean (Hz)	19.5	19.2
power max (dB)	3.12	3.02
power mean (dB)	3.16	3.09

Table 4 Prediction result of prosodic features of reactive tokens {*a:*, *ha:*} (square root of MSE)

	baseline	prediction
F0 max (Hz)	32.4	28.3
F0 mean (Hz)	33.0	29.6
power max (dB)	4.08	3.84
power mean (dB)	4.26	4.09

6 Conclusions

We have investigated whether and how synchrony is expressed by the prosodic features of backchannels with respect to the preceding speaker's utterances. To this objective, we recorded counseling sessions, in which a counselor conducts attentive listening by generating backchannels frequently and carefully.

The major findings in this work are summarized as follows:

- There is a different tendency between acknowledging tokens and reactive tokens. The reactive tokens are more likely to have synchrony.
- In addition to F0, the power feature plays an important role. Specifically, the power feature tends to have more correlation patterns for repeated tokens and reactive tokens.

Based on the observations, we have tried prediction of the prosodic features, given those of the preceding speaker's utterances. The model suggests possibility of appropriately controlling the prosodic parameters of backchannels generated by a conversational system to make it more natural and friendly to users.

Future work includes development of a more elaborate model which incorporates rich contextual information and also is capable of predicting timing and segmental patterns of the backchannels.

Acknowledgement: This work was supported by JST CREST and ERATO.

References

1. N.Ward, D.Novick, L.P.Morency, T.Kawahara, D.Heylen, and J.Edlund, editors. *Proc. Interdisciplinary Workshop on Feedback Behaviors in Dialog*, 2012.
2. N.Ward. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody*, pages 325–328, 2004.

3. T.Kawahara, Z.Q.Chang, and K.Takanashi. Analysis on prosodic features of Japanese reactive tokens in poster conversations. In *Proc. Int'l Conf. Speech Prosody*, 2010.
4. R.Levitan and J.Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proc. InterSpeech*, pages 3081–3085, 2011.
5. B.Xiao, P.G.Georgiou, Z.E.Imel, D.C.Atkins, and S.Narayanan. Modeling Therapist Empathy and Vocal Entrainment in Drug Addiction Counseling. In *Proc. Interspeech*, pages 2861–2865, 2013.
6. N.Ward. Using prosodic clues to decide when to produce back-channel utterances. In *Proc. ICSLP*, pages 1728–1731, 1996.
7. N.Ward and W.Tsukahara. Prosodic features which cue back-channel responses in English and Japanese. *J. Pragmatics*, 32(8):1177–1207, 2000.
8. 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 dialogs. *Language & Speech*, 41(3-4):295–321, 1998.
9. H.Noguchi and Y.Den. Prosody-Based Detection of the Context of Backchannel Responses. In *Proc. ICSLP*, volume 2, pages 8570–8573, 1998.
10. T.Solorio, O.Fuentes, N.G.Ward, and Y.A.Bayyari. Prosodic Feature Generation for Back-Channel Prediction. In *Proc. InterSpeech*, pages 2398–2401, 2006.
11. A.Gravano and J.Hirschberg. Backchannel-Inviting Cues in Task-Oriented Dialogue. In *Proc. InterSpeech*, pages 1019–1022, 2009.
12. K.P.Truong, R.Poppe, and D.Heylen. A Rule-Based Backchannel Prediction Model Using Pitch and Pause Information. In *Proc. InterSpeech*, pages 3058–3061, 2010.
13. N.Kitaoka, M.Takeuchi, R.Nishimura, and S.Nakagawa. Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems. *J. Japanese Society for Artificial Intelligence*, 20(3):220–228, 2005.
14. S.Fujie, K.Fukushima, and T.Kobayashi. Back-Channel Feedback Generation Using Linguistic and Nonlinguistic Information and its Application to Spoken Dialogue System. In *Proc. InterSpeech*, pages 889–892, 2005.
15. Y.Kamiya, T.Ohno, and S.Matsubara. Coherent back-channel feedback tagging of in-car spoken dialogue corpus. In *Proc. SIGdial*, 2010.
16. D.Ozkan and L.-P.Morency. Modeling wisdom of crowds using latent mixture of discriminative experts. In *Proc. ACL/HLT*, 2011.
17. T.Kawahara, S.Hayashi, and K.Takanashi. Estimation of interest and comprehension level of audience through multi-modal behaviors in poster conversations. In *Proc. INTERSPEECH*, pages 1882–1885, 2013.
18. M.Heldner, J.Edlund, and J.Hirschberg. Pitch similarity in the vicinity of backchannels. In *Proc. InterSpeech*, pages 3054–3057, 2010.
19. K.Maekawa. Corpus of Spontaneous Japanese: Its design and evaluation. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 7–12, 2003.