Synchrony in Prosodic and Linguistic Features between Backchannels and Preceding Utterances in Attentive Listening

Tatsuya Kawahara, Takashi Yamaguchi, Miki Uesato, Koichiro Yoshino, Katsuya Takanashi
School of Informatics, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

***Abstract*—In human-human dialogue, especially in attentive listening such as counseling, backchannels play an important role. Appropriately coordinated backchannels will not only make smooth communication but also help establish rapport. By collecting counseling dialogue, we investigate whether and how synchrony is expressed by prosodic and linguistic features of backchannels with respect to the preceding speaker's utterances. First, 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. Next, we investigate the relationship between the morphological complexity of backchannels and the syntactic complexity of the preceding clause/sentence unit. The result can be useful for generating a variety of backchannels adaptive to the speaker's utterances.**

Index Terms: dialogue, backchannel, prosody

I. Introduction

In recent years, a number of spoken dialogue systems have been deployed in smart phones and car navigation systems to conduct simple tasks and information retrieval. These systems basically assume that a user makes one utterance per one turn, which is responded by the system, and the utterances are not supposed to overlap each other. This "half duplex" communication mode is much different from that of human-human dialogue, in which one person often has a long speech turn consisting of many utterances while the listener gives an occasional feedback within the turn. 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. 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 suggest that the listener is listening, understanding, and agreeing to the speaker.

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 entrainment [2]. 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 [3]. The work presented in this paper focuses on the synchrony effect of backchannels rather than their individual role and meaning. Specifically, we first investigate whether prosodic synchrony is observed in generating backchannels in counseling dialogue by analyzing the relationship between prosodic features of the backchannels and those of preceding utterances by the speaker. Next, we investigate another synchrony effect by analyzing the relationship between the morphological structure of backchannels and the syntactic structure of the preceding utterances.

The finding of this work would be useful for designing a new kind of spoken dialogue systems or conversational agents, which conduct attentive listening to particular user populations such as elderly and ill persons. The current spoken dialogue systems usually generate the same or limited patterns of backchannels in terms of lexical choice and prosodic parameters. However, in order to make smooth communication by establishing rapport, it is very important to generate appropriate backchannels adaptive to the speaker's utterances. Note that speech recognition and understanding may not be necessary to realize this function.

II. 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 acknowledgment 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 the following three factors: timing (when), prosody (how) and lexical entry (what).

A. Timing

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. [4], [5] pointed out the low pitch as a major prosodic cue of backchannels. Koiso et al. [6] introduced a decision tree to derive rules from prosodic and morphological patterns.

There are also several studies which actually implemented a dialogue system to generate backchannels using a decision

tree [7]. Recently, more elaborate discriminative modeling and an efficient learning mechanism using the wisdom of crowds have been introduced [8], [9].

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

B. Lexical entry

The lexicon of backchannels is language-dependent. In general, morphological patterns of backchannels are classified into two categories. One is lexical tokens, and they are usually same entries as acknowledging tokens such as "okay" and "right" in English and "*hai*" and "*un"* in Japanese. In Japanese, their lexical patterns are limited, but can be repeated, for example, "*un un un*". The other category is reactive tokens, which are often non-lexical, such as "wow" in English and "*he:*" in Japanese. The acknowledging tokens are more frequently used and they indicate that the listener is listening and understanding, while the reactive tokens are specially used to indicate the listener's strong reaction and assessment to what is uttered.

Although the role of individual tokens has been discussed in relation with the listener's state of mind, there is almost no previous work on the choice of lexical patterns according to the context of the preceding utterance by the speaker.

C. Prosody

Prosody of backchannels is important especially for expressing assessment, and we have identified particular patterns to express interest and surprise in conversations [10], [11]. 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 "okay" or "*hai*", which gives a monotonic impression.

III. 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 career path, to a counselor. The subjects are eight 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 four sessions. All participants are Japanese native.

The dialogue started with some chatting and the following counseling session lasted around 20-30 minutes. It is observed that the counselors make backchannels every 5 to 7 seconds, that is at almost every end of the speaker's utterances.

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 (perceptually) and the number of the repetitions, and represented, for example, by "*un* x2" or "(*un*)+".

IV. Analysis on Prosodic Synchrony between Backchannels and Preceding Utterances

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

A. 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.

B. General Synchrony in Prosodic Features

We first investigate the general tendency of synchrony, which was reported by Heldner et al. [12]. 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. [12] 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 (p-value< 0.05) 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).

Although we do not have a clear explanation for the results, the difference in general prosodic patterns between English

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

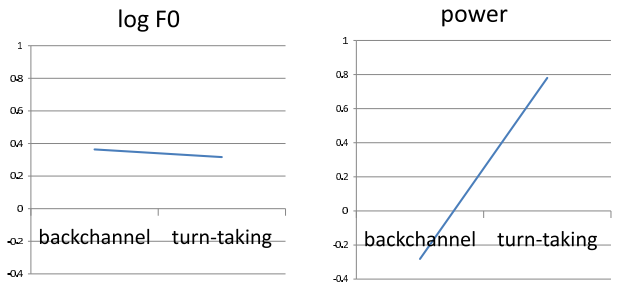

Fig. 1. Difference in prosodic features from the preceding utterances (comparison of backchannels and normal turn-taking)

TABLE I
CORRELATION OF PROSODIC FEATURES BETWEEN BACKCHANNELS AND THE PRECEDING UTTERANCES

morph. pattern	count	log F0	power
u:n	225	max (0.14)	max (0.14) mean (0.18)
un	361	max (0.22) mean (0.12)	max (0.23)
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)
a:	28	mean (0.22)	mean (0.25)
ha:	27	mean (0.23)	max (0.47) mean (0.29)

and Japanese and also the difference in segmental and prosodic patterns of backchannels may be attributed.

C. 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 morphological pattern, but those with fewer occurrence counts (less than 25) are not used.

We compute F0 and power in the same manner as in the previous sub-section, and parameterize them with their mean, maximum (max) and range within the segment. Then, a correlation coefficient is computed between the parameter of backchannels and that of the preceding utterances.

The results of the correlation analysis are presented in Table I. Here we list those with significant correlation (p-value<0.05) for the acknowledging tokens (upper part of Table I) and those larger than 0.20 for the reactive tokens (lower part of Table I) 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. 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.

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

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

V. Analysis on Morphological Patterns of Backchannels and Syntactic Features of Preceding Utterances

Next, we investigate the relationship between morphological patterns of backchannels and syntactic features of preceding utterances. Here, we hypothesize that the lexical choice of backchannels is affected by the syntactic patterns of the preceding utterances, and that, especially in Japanese, the repetition patterns are affected by the linguistic complexity of the preceding utterances; more complex morphological patterns are used when the preceding utterance is long, complex or at the end of complete sentences. This realizes a rhythmic effect with respect to the linguistic features.

The utterances (IPUs) are chunked into clause and sentence units according to the guideline of the CSJ [13]. The boundaries are annotated with three types: sentence boundary, strong clause boundary, and weak clause boundary. While strong clause boundaries usually appear between parallel clauses, weak clause boundaries are defined before/after the clauses that depend on other clauses, such as those starting with "because".

A. Relationship between Boundary Type and Morphological Patterns of Backchannels

The frequency distribution of the backchannel clusters according to the boundary type is plotted in Figure 2. Here, "*un* x4" is merged into "*un* x3".

A clearly different tendency is observed between the sentence boundary and the two clause boundary types. While simple morphological patterns of backchannels, mostly one or two repetitions of acknowledging tokens, "*un*" and "*un un*", are more often used in the clause boundaries, the sentence boundaries are more likely to be followed by reactive tokens. Simple patterns are preferred to encourage the speaker to keep talking, and reaction with prominent prosodic patterns, as discussed in the previous section, are expressed after the speaker completes a sentence. This suggests that stronger reactions are made in stronger boundaries. However, there is no clear distinction between the weak clause boundary and the strong clause boundary. It is also infeasible to control the number of repetitions with this feature.

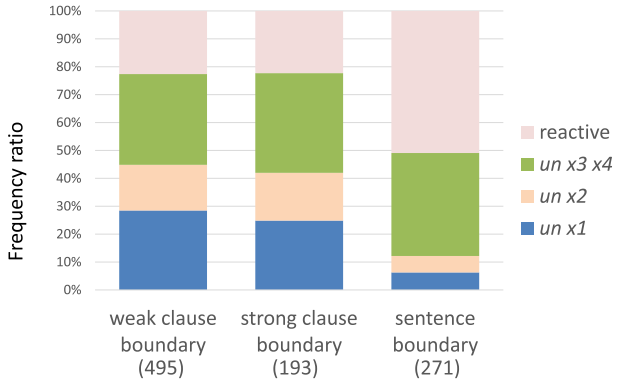

Fig. 2. Frequency distribution of backchannel clusters according to the boundary type of the preceding utterances.

TABLE II
STATISTICS OF SYNTACTIC FEATURES IN THE PRECEDING CLAUSE/SENTENCE UNIT

	un	*un* x2	*un* x3 x4	reactive
no. of phrases	**4.73**	**5.52**	5.42	5.15
depth of parse tree	**2.18**	**2.57**	2.56	2.54
width of parse tree	1.88	2.00	1.89	1.75

B. Relationship between Syntactic Complexity and Morphological Patterns of Backchannels

We also conduct a morphological and syntactic analysis on the preceding clause/sentence unit and investigate the relationship between the linguistic features and the morphological patterns of the backchannels. Specifically, we count the number of *bunsetsu* phrase units and the depth and the width of the parse tree generated by a Japanese syntactic parser KNP. [2] The width means how many phrases depend on the verb, and the depth means how many dependencies at maximum exist before the verb. These measures show the complexity of the clause/sentence unit.

The statistics according to the backchannel clusters are shown in Table II. The difference in the first two measures between "*un*" and "*un* x2", shown in bold fonts, are statistically significant (p-value$<$ 0.05). Interestingly, these two short acknowledging tokens cannot be distinguished in the analysis in the previous subsection, but it is reasonable that repeated tokens are used when the number of phrases is larger or the depth of the parse tree is deeper, which involves more components in the unit. This is regarded as a synchrony with regard to the linguistic patterns.

Overall, the hypothesis on the relationship between the complexity of the preceding utterances and the complexity of the morphological patterns of the backchannels is verified to some extent.

VI. CONCLUSIONS

We have investigated whether and how synchrony is expressed by prosodic and linguistic 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 tend to have more correlation patterns for repeated tokens and reactive tokens.
- The morphological complexity of backchannels is also related with the syntactic complexity of the preceding clause/sentence unit.

The findings will be useful for appropriately determining the lexical choice and the prosodic parameters of backchannels generated by a conversational system to make it more natural and friendly to users. For example, when the speaker utters a longer clause, a backchannel should be repeated, and when a sentence is completed, a reactive token should be made, with power adjusted to the speaker's voice. We plan to develop this kind of system and evaluate with human subjects.

Acknowledgment: This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction program.

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] R.Levitan and J.Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proc. InterSpeech*, pages 3081–3085, 2011.
[3] B.Xiao, P.G.Georgiou, Z.E.Imel, D.Atkins, and S.Narayanan. Modeling therapist empathy and vocal entrainment in drug addiction counseling. In *Proc. InterSpeech*, pages 2861–2864, 2013.
[4] N.Ward. Using prosodic clues to decide when to produce back-channel utterances. In *Proc. ICSLP*, pages 1728–1731, 1996.
[5] N.Ward and W.Tsukahara. Prosodic features which cue back-channel responses in English and Japanese. *J. Pragmatics*, 32(8):1177–1207, 2000.
[6] 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.
[7] 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.
[8] Y.Kamiya, T.Ohno, and S.Matsubara. Coherent back-channel feedback tagging of in-car spoken dialogue corpus. In *Proc. SIGdial*, 2010.
[9] D.Ozkan and L.-P.Morency. Modeling wisdom of crowds using latent mixture of discriminative experts. In *Proc. ACL/HLT*, 2011.
[10] 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.
[11] 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.
[12] M.Heldner, J.Edlund, and J.Hirschberg. Pitch similarity in the vicinity of backchannels. In *Proc. InterSpeech*, pages 3054–3057, 2010.
[13] 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.

[2] `http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP`