INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Prediction and Generation of Backchannel Form for Attentive Listening Systems

Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel Ward[*]

Kyoto University, School of Informatics, Sakyo-ku, Kyoto 606-8501, Japan
[*]University of Texas at El Paso, El Paso, Texas 79968, USA

Abstract

In human-human dialogue, especially in attentive listening such as counseling, backchannels are important not only for smooth communication but also for establishing rapport. Despite several studies on when to backchannel, most of the current spoken dialogue systems generate the same pattern of backchannels, giving monotonous impressions to users. In this work, we investigate generation of a variety of backchannel forms according to the dialogue context. We first show the feasibility of choosing appropriate backchannel forms based on machine learning, and the synergy of using linguistic and prosodic features. For generation of backchannels, a framework based on a set of binary classifiers is adopted to effectively make a "not-to-generate" decision. The proposed model achieved better prediction accuracy than a baseline which always outputs the same backchannel form and another baseline which randomly generates backchannels. Finally, evaluations by human subjects demonstrate that the proposed method generates backchannels as naturally as human choices, giving impressions of understanding and empathy.
Index Terms: spoken dialogue systems, attentive listening, backchannel

1. Introduction

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 immediately responded to by the system, and the utterances are not supposed to overlap. This "half duplex" communication mode is very different from that of human-human dialogue, in which one person often has a long speech turn consisting of many utterances, during which the listener gives occasional feedback. 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 signal that the listener is listening and understanding, while suggesting that the current speaker can keep the dialogue turn. Backchannels are also used to express the listener's reaction or assessment such as surprise, interest and empathy. Without the backchannel feedback, the speaker would be anxious whether the communication is well maintained and would feel as if talking to a dumb "machine". In addition to the effect of individual backchannels, backchanneling can also contribute to a sense of rhythm, to entrainment, and to feelings of synchrony, contingency, and rapport [2, 3]. In counseling dialogues in particular, it is crucial for a counselor to keep the client talking, and one technique they use is effective backchanneling to express empathy and create synchrony [4].

Computational models of backchanneling have mostly addressed the question of when to backchannel, including investigating the role on low-pitch regions and other prosodic and lexical features as cues for backchannels [5, 6, 7, 8, 9, 10, 11], and developing more effective discriminative modeling and more efficient learning mechanisms [12, 13]. Some dialogue systems have applied this to generate backchannels, some producing very effective attentive listening behavior [14, 15, 16].

However, current systems do not necessarily vary the backchannels, in terms of morphological form and prosody, often giving monotonous impressions to users. In previous work [17], we investigated the correlations or synchrony effect in the prosody of backchannels.

In this work, we focus on the morphological (lexical) form of backchannels. We first analyze how the choice of the backchannel form is related with the linguistic features of the preceding utterances of the speaker. Then, we attempt to predict the backchannel form, given the context of the preceding utterance. Specifically, machine learning is conducted using the linguistic and prosodic features in the context. Finally, we realize automatic generation of backchannels dependent on the dialogue context. It is evaluated by human subjects in comparison with random generation and the actual counselor's choice. Although the role of individual forms has been discussed in relation with the listener's state of mind or intended pragmatic effect in conversation analysis and linguistic studies [18, 19, 20, 21, 22, 23, 24, 25], there is almost no previous work on the choice of backchannel form according to the context of the preceding utterance by the speaker.

In the remainder of the paper, after a description of the counseling corpus and the annotation of backchannels in Section 2, analysis, prediction and generation of the backchannel form using the corpus are addressed in Section 3, 4 and 5, respectively. Finally, a subjective evaluation of generated backchannels is presented in Section 6, before the conclusion in Section 7.

2. Corpus of Counseling Dialogue

2.1. Dialogue Data Collection and Transcription

We 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, regarding human relationships and career paths, with a counselor. The subjects were eight college students, 20 to 25 years old. We had two counselors (a male with 7-years of counseling experience and a female with 4-years), and each took part in four sessions. All participants were native speakers of Japanese. Each dialogue started with some chatting, and lasted

http://dx.doi.org/10.21437/Interspeech.2016-118

Table 1: Categories and occurrence counts of backchannels

category	occurrence counts
un	319 (154)
*un*x2	183 (130)
*un*x3	357 (284)
assessments	215 (209)
none	1601 (685)
total	2675 (1462)

counts in parentheses are those observed at clause boundaries

around 20-30 minutes.

Speech was captured by a head-set microphone worn by each participant and transcribed following the guidelines of the Corpus of Spontaneous Japanese (CSJ) [26]. This includes annotation of IPUs (InterPausal Unit) and linguistic clause units.

2.2. Categories of Backchannel Forms

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 often referred to as "continuers", and distinguished from acknowledgment and fillers, which are used to take or keep a turn.

Many backchannels share the same morphological form as acknowledging tokens such as "okay" and "right" in English and "*hai*" and "*un*" in Japanese. In Japanese, the possible backchannels are fairly limited lexically, but these can be repeated, as in "*un un un*" Since it is difficult to acoustically or contextually distinguish "*un*" from "*hun*", we treat them together. We group all such forms into three categories, based on the number of the repetitions, and represented, for example, by "*un*x2" and "*un*x3".

In addition, there are other backchannel types commonly used to express the listener's reaction or assessment such as surprise, interest and empathy. Their morphological forms are often non-lexical, such as "wow" in English and "*he:*" in Japanese. We treat these as the "assessments" category.

Backchannels were annotated as such in the counseling corpus. Table 1 summarizes the categories of the backchannel forms dealt with in this work and their occurrence counts in the corpus. Here, repetitions of more than three times are merged into the *un*x3 category. In the table, "none" is the number of IPUs and clause boundaries not followed by any backchannel. Forms not included in the four categories are excluded from this study, as they are infrequent and often difficult to annotate.

It is observed that the counselors produce backchannels very frequently, at approximately 40% of IPU boundaries. Most of them are continuers, with the different forms (the varying numbers of repetitions) used in a good balance. Assessments are not so frequent, but they are more prominent than continuers when they do occur.

2.3. Additional Annotation of Backchannels

Obviously, generation of backchannels and the choice of their form are arbitrary to some extent, although not anything is possible. The choice is also person-dependent. The statistical model for prediction and generation of backchannels should deal with these problems of variety by using a large-scale corpus. But for evaluation there remains the problem of arbitrary variation: evaluating a model on its ability to predict exactly the form observed in the corpus would understate its actual pragmatic ability.

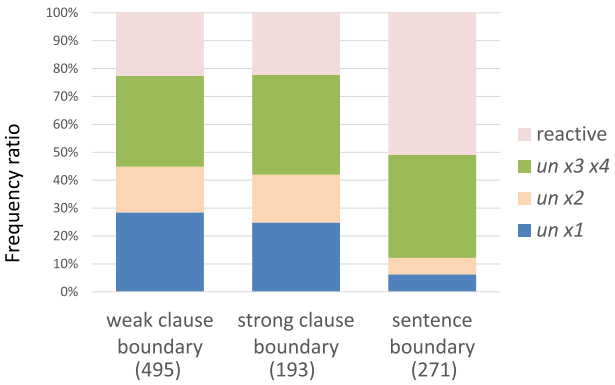

Figure 1: Frequency distribution of backchannel form categories according to the boundary type of the preceding utterances

To deal with this, we augmented the annotation of backchannels for evaluation purposes. Specifically, we had three human annotators judge which backchannel form categories are acceptable for a given prediction point (=end of IPU or clause) after listening to the preceding dialogue context. We consider a category acceptable in a position if all three annotators chose it as a possible label there.

3. Analysis of Backchannel Forms and Linguistic Features of Preceding Utterance

First, we investigate the relationship between the morphological form of backchannels and the linguistic features of preceding utterances. The analysis in this section is limited to the backchannels observed at clause boundaries in the corpus.

3.1. Relationship between Boundary Type and Backchannel Form

Each clause boundary is annotated as one of 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 dependent clauses, such as those starting with "because". The frequency distribution of the backchannel form categories versus the boundary type is plotted in Figure 1.

Clearly different backchannel occurrence tendencies are observed for sentence boundaries versus the two clause boundary types. While simple continuers ("*un*" and "*un*x2") are more often used in the clause boundaries, the sentence boundaries are more likely to be followed by assessments. Continuers are preferred to encourage the speaker to keep talking, and assessments are expressed mostly when the speaker completes a sentence. However, there is no clear difference between the weak clause boundary and the strong clause boundary, and boundary types do not help determine the number of repetitions.

3.2. Relationship between Syntactic Complexity and Backchannel Form

We also conduct a morphological and syntactic analysis of the preceding clause/sentence unit and investigate the relationship

Table 2: Statistics of syntactic features in the preceding clause/sentence unit

	un	*un*x2	*un*x3	assessments
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

between its linguistic features and the morphological form of the backchannels. Specifically, we count the number of *bunsetsu* phrase units and the depth and the width of the parse tree generated by the Japanese syntactic parser KNP, where the width is the number of phrases depending on the verb, and the depth is the maximum number of dependencies before the verb. These are measures of the complexity of the clause/sentence unit.

The statistics for the backchannel form categories are shown in Table 2. The differences in the first two measures between "*un*" and "*un*x2", shown in bold, are statistically significant. Although these two continuers could not be distinguished in the previous subsection, they are separable by considering these syntactic complexity measures. It is shown that repeated tokens are more likely to be used when the number of phrases is larger.

4. Selection of Backchannel Form

Based on this analysis, we design a feature set used for selection of the backchannel form category. In addition to the features mentioned in the previous section, the ending word and its POS (part of speech) are added. We also incorporate history information (up to two previous) of the clause boundary type and of the previous backchannel form categories.

In addition, as previous work has shown that prosodic information is useful for predicting occurrence of backchannels (although prediction of the backchannel form has not been done before), we add to the above the following prosodic features: duration, $\Delta \log$ F0, range of $\log$ F0, Δ power, range of power, and creakiness. Except for duration, these are computed over the final 150 msec of the preceding utterance.

Using these features, a logistic regression model is trained for selection of the backchannel form categories. In this section, the problem is formulated as classification of the four categories, given a clause boundary in which any backchannel is observed in the corpus. The model is trained and evaluated in a cross-validation manner in which one session is held out for evaluation. The augmented annotation described in Section 2.3 is used in this evaluation: we regard a prediction as correct if it matches the original form in the corpus or any of the annotator-added forms.

For reference, we use a baseline where the most frequent form (*un*x3) is always output. We also test another baseline method which randomly generates backchannels according to the observed frequency distribution of the four categories. Average performance over 1000 trials is computed.

The classification accuracy by these methods is listed in Table 3. The model with only the linguistic features significantly outperformed both baseline methods. Incorporation of the prosodic features further improved the performance. The results confirm that the proposed model using both feature sets works effectively.

The detailed performance (recall, precision and F-measure) by the proposed model is shown for each backchannel category in Table 4. It is observed that the precision of simple continuers (*un* and *un*x2) is very high because they are acceptable in many cases. The precision of assessment tokens is not so high, but still much better than the random baseline (0.389), which is critical for good subjective impressions.

Table 3: Classification performance of backchannel form category given a clause boundary in which any backchannel is observed

method	accuracy
baseline: most frequent	41.7%
baseline: weighted random	53.5%
prosodic only	53.1%
linguistic only	62.2%
linguistic + prosodic	65.6%

Table 4: Selection performance per backchannel form category by the proposed model using linguistic and prosodic features

category	recall	precision	F-measure
un	0.460	0.982	0.626
*un*x2	0.484	0.826	0.611
*un*x3	0.830	0.488	0.615
assessments	0.703	0.654	0.678
average	0.656	0.656	0.656

5. Flexible Prediction of Backchannels

Next, we extend the model so that it can also determine right after each user utterance (IPU), whether or not to generate a backchannel there. This can be formulated as a five-class classification problem by adding the category "not to generate any backchannel" to the four categories of the backchannel form.

We also adopt a different model, which predicts using a set of binary classifiers. Here we train a binary classifier for each backchannel form category, which computes the probability the generation of that category there. The final action is determined by combining the results of the four binary classifiers. If none of them exceeds a threshold, the system does not generate any backchannel. If one or more classifiers have an output exceeding the threshold, the system selects the one with the highest score. In this work, the threshold was tuned so that the number of predicted backchannels is similar to the observed count in the corpus. At times when this threshold (0.275) was exceed by any of the binary classifiers, it was often exceeded by several of them (2.3 on average).

The five-class classifier and the combined classifier are both trained using the same logistic regression model and the same feature set described in the previous section, but the training and evaluation samples are chosen by IPU instead of the clause boundaries.

The prediction accuracy by cross-validation is shown in Table 5, with comparison to the weighted random generation method. Compared with Table 3, the random generation method significantly drops in performance because more than half the samples are "not-to-generate". In contrast, the proposed method can generally determine whether to generate a backchannel, maintaining the overall performance. Among the proposed models, the combination of four binary classifiers is more effective than the five-class classifier. The training of a single

Table 7: Subjective evaluation of generated backchannels

item	weighted random	proposed	counselor
Q1: Are backchannels natural in general?	-0.42	1.04**	0.79
Q2: Are backchannels in good tempo in general?	0.25	1.29*	1.00
Q3: Did you feel the system listened with care?	0.33	1.25	0.96
Q4: Did you feel the system understood well?	-0.13	1.17**	0.79
Q5: Did you feel the system showed interest?	0.21	1.21	1.04
Q6: Did you feel the system showed empathy?	0.13	1.04*	0.46
Q7: Would you like to talk to this system?	-0.33	0.96**	0.29

$*p < 0.05$, $**p < 0.01$ compared against "weighted random" baseline

Table 5: Prediction performance of backchannel generation and form category choice given an IPU boundary

method	accuracy
weighted random	43.1%
5-class classifier	61.0%
set of 4 binary classifiers	64.3%

Table 6: Prediction performance per backchannel form category by the proposed model using the set of 4 binary classifiers

category	recall	precision	F-measure
un	0.311	0.657	0.422
*un*x2	0.382	0.820	0.521
*un*x3	0.672	0.333	0.454
assessments	0.467	0.342	0.405
not-to-generate	0.775	0.769	0.772
average	0.643	0.643	0.643

five-class classifier is difficult because of the large proportion and different nature of the category of "not-to-generate".

The detailed prediction performance (recall, precision and F-measure) of the combination of four binary classifiers is shown for each backchannel category in Table 6. Although prediction accuracy for the backchannel form is degraded from Table 4, both recall and precision of "not-to-generate" is very high, meaning that the decision of whether or not to generate is itself reliable.

6. Subjective Evaluation of Generated Backchannels

Finally, we conduct a subjective evaluation of backchannels generated by the proposed method. Based on the results in the previous section, the combination of four binary classifiers is adopted. A set of voice files for the backchannel forms were prepared by a voice actress. There are variants in prosodic expression for each form, but we chose one pattern (file) for each form in this experiment. The audio channels of the counselors in the counseling dialogues described in Section 2 were replaced by the system-generated backchannels.

For comparison, we prepared similar simulated-dialogue audio files using the weighted random backchannel generation and also by using the original backchannel forms by the counselors. In the latter case, the voice was replaced by the actress-produced samples to enable a fair comparison based on only the selection of the backchannel form.

Nine subjects, 5 male and 4 female, listened to eight segments of approximately two minutes extracted from the audio files. Then they filled in a questionnaire of several items regarding their impression of the system on a seven-point scale, from -3, "totally disagree", to 3, "totally agree". The average evaluation results on key items are given in Table 7, comparing the proposed method with the weighted random generation method and the counselors' choices.

It is observed that the proposed generation method obtained a higher rating than the random generation method in all items. Statistically significant differences are marked with an asterisk in the table. The random generation method was rated particularly low in Q1 regarding the naturalness of the backchannels, leading to a very negative result in Q7. This result confirms that backchannels should be generated appropriately depending on the dialogue context. In more detail, there is not a significant difference in Q3 ("system listening") and Q5 ("showing interest"), but the proposed system obtained a higher rating in Q4 ("system understanding") and Q6 ("showing empathy").

We note that the results of the proposed method are comparable to those for the counselors' choices: in no item is there a statistically significant difference between them. Indeed, the ratings of the counselors' choices are not so high. This is probably because the same voice file was used for each backchannel form, with no variation in prosody. This suggests the need to change the prosody of backchannels according to the dialogue context. We plan to incorporate our previous finding on the synchrony effect of the prosody of the backchannels with the preceding utterances [17]. Moreover, there may be also a need to tune the precise timing of backchannels, that is the interval from the end of the preceding user utterance.

7. Conclusions

We have proposed a model that generates a variety of backchannel forms appropriate for the dialogue context. First, we showed that different kinds of backchannel forms are used depending on the boundary type and the number of phrases in the preceding utterance. Then, we conducted machine learning using these features in combination with prosodic features. This model achieved a classification accuracy of 66%, significantly outperforming the most-frequent-class baseline and the random generation method. We also designed a two-step framework which combines binary classifiers, each designed for an individual backchannel form category. This is very effective for generation of backchannels including "not-to-generate" decision. The subjective evaluation demonstrates that the proposed method generates backchannels more naturally, giving impressions of understanding and empathy.

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

8. References

[1] N.Ward, D.Novick, L.P.Morency, T.Kawahara, D.Heylen, and J.Edlund, Eds., *Proc. Interdisciplinary Workshop on Feedback Behaviors in Dialog*, 2012.

[2] J.Gratch, A.Okhmatovskaia, F.Lamothe, S.Marsella, M.Morales, R. der Werf, and L.-P.Morency, "Virtual rapport," in *Proc. Intelligent Virtual Agents*, 2006, pp. 14–27.

[3] R.Levitan and J.Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proc. InterSpeech*, 2011, pp. 3081–3085.

[4] 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*, 2013, pp. 2861–2864.

[5] N.Ward, "Using prosodic clues to decide when to produce backchannel utterances," in *Proc. ICSLP*, 1996, pp. 1728–1731.

[6] N.Ward and W.Tsukahara, "Prosodic features which cue backchannel responses in English and Japanese," *J. Pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[7] 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*, vol. 41, no. 3-4, pp. 295–321, 1998.

[8] T.Solorio, O.Fuentes, N.G.Ward, and Y.Al Bayyari, "Prosodic Feature Generation for Back-Channel Prediction," in *Proc. InterSpeech*, 2006, pp. 2398–2401.

[9] A.Gravano and J.Hirschberg, "Backchannel-Inviting Cues in Task-Oriented Dialogue," in *Proc. InterSpeech*, 2009, pp. 1019–1022.

[10] K.P.Truong, R.Poppe, and D.Heylen, "A Rule-Based Backchannel Prediction Model Using Pitch and Pause Information," in *Proc. InterSpeech*, 2010, pp. 3058–3061.

[11] I. de Kok and D.Heylen, "Integrating backchannel prediction models into embodied conversational agents," in *Intelligent Virtual Agents*, 2012, pp. 268–274.

[12] Y.Kamiya, T.Ohno, and S.Matsubara, "Coherent back-channel feedback tagging of in-car spoken dialogue corpus," in *Proc. SIGdial*, 2010.

[13] D.Ozkan and L.-P.Morency, "Modeling wisdom of crowds using latent mixture of discriminative experts," in *Proc. ACL/HLT*, 2011.

[14] 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*, vol. 20, no. 3, pp. 220–228, 2005.

[15] M.Schröder, E.Bevacqua, R.Cowie, F.Eyben, H.Gunes, D.Heylen, M. Maat, S.Pammi, M.Pantic, C.Pelachaud, B.Schuller, E. Sevin, M.Valstar, and M.Wollmer, "Building autonomous sensitive artificial listeners," *IEEE Trans. Affective Computing*, vol. 3, pp. 165–183, 2012.

[16] D.DeVault, R.Artstein, G.Benn, T.Dey, E.Fast, A.Gainer, K.Georgila, J.Gratch, A.Hartholt, M.Lhommet, G.Lucas, S.Marsella, F.Morbini, A.Nazarian, S.Scherer, G.Stratou, A.Suri, D.Traum, R.Wood, Y.Xu, A.Rizzo, and L-P.Morency, "SimSensei Kiosk: A virtual human interviewer for healthcare decision support," in *Proc. AAMAS*, 2014.

[17] T.Kawahara, M.Uesato, K.Yoshino, and K.Takanashi, "Toward adaptive generation of backchannels for attentive listening agents," in *Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS)*, 2015.

[18] K.Drummond and R.Hopper, "Back channels revisited: Acknowledgment tokens and speakership incipiency," *Research on Language and Social Interaction*, vol. 26, pp. 157–177, 1993.

[19] X.Deng, "The use of listener responses in Mandarin Chinese and Australian English conversations," *Pragmatics*, vol. 18, pp. 303–328, 2008.

[20] R.Gardner, *When Listeners Talk: Response tokens and listener stance*. John Benjamins, 2001.

[21] Y.Katagiri, M.Sugito, and Y.Nagano-Madsen, "Forms and prosodic characteristics of backchannels in tokyo and osaka japanese," *Proc. ICPhS*, pp. 2411–2414, 1999.

[22] N.Ward, "Non-lexical conversational sounds in American English," *Pragmatics and Cognition*, vol. 14, pp. 113–184, 2006.

[23] A.Golato and Z.Fagyal, "Comparing single and double sayings of the german response token ja and the role of prosody: A conversation analytic perspective," *Research on Language and Social Interaction*, vol. 41, pp. 241–270, 2008.

[24] T.Stivers, "'No no no' and other types of multiple sayings in social interaction," *Human Communication Research*, vol. 30, pp. 260–293, 2004.

[25] D.Wong and P.Peters, "A study of backchannels in regional varieties of English, using corpus mark-up as the means of identification," *Int'l J. Corpus Linguistics*, vol. 12, pp. 479–509, 2007.

[26] K.Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 7–12.