A rule-based backchannel prediction model using pitch and pause information

Khiet P. Truong, Ronald Poppe, and Dirk Heylen

Human Media Interaction Group, University of Twente, The Netherlands
{k.p.truong,poppe@ewi.utwente.nl,d.k.j.heylen}@ewi.utwente.nl

Abstract

We manually designed rules for a backchannel (BC) prediction model based on pitch and pause information. In short, the model predicts a BC when there is a pause of a certain length that is preceded by a falling or rising pitch. This model was validated against the Dutch IFADV Corpus in a corpus-based evaluation method. The results showed that our model performs slightly better than another well-known rule-based BC prediction model that uses only pitch information. We observed that the length of a pause preceding a BC is one of the important features in this model, next to the duration of the pitch slope at the end of an utterance. Further, we discuss implications of a corpus-based approach to BC prediction evaluation.

Index Terms: backchannel prediction, backchannel, prosody, rule-based

10.21437/Interspeech.2010-59

1. Introduction

During conversations where there is a listener and a speaker, it is very natural for the listener to send short signals of understanding and attentiveness to the speaker to show that he/she is following the conversation. Not doing so would result in 'odd' and disruptive conversations. These short signals are also known as *backchannel feedback* or *backchannels* (BCs) and are comprised of various types of vocal and visual behaviour such as head nods and vocalizations, e.g. 'yeah' or 'mm-hmm'. The term *backchannel* was first introduced by Yngve [1] who described the *back channel* (as opposed to the *main* channel) as a communicative channel over which these short signals are sent without relinquishing the turn. A lot of attention has been paid by researchers to the precise timing and the context in which BCs occur. It is said that speakers display certain speaker signals that may cue backchannel opportunities (e.g. [2]). Dittmann and Llewellynn [3] found that most backchannel responses occur at the end of clauses which sounded most 'final'. Both pause length and juncture type (falling, rising or sustained intonation) were significantly related to listener responses. Ward and Tsukahara [4] consider a region of low pitch as an important cue for backchannel placement. Furthermore, rising or falling final intonations have also been found to cue backchannels [5, 6, 7]. Based on these findings, prosodic, pause, and lexical features have been used by other researchers to train statistical and rule-based models of BC prediction, e.g. [4, 8, 9, 10].

In the context of Embodied Conversational Agents (ECAs), our goal is to develop an automatic rule-based BC prediction model that can predict BCs in an online setting using pause information, and the rising and falling intonation property as suggested in several studies [5, 8, 7, 6]. In a similar fashion as Ward and Tsukahara [4], we manually designed rules that determine when to backchannel. Our rules were tested in a previous experiment [11] on a small dataset; the results were very promising, not only performance-wise, but also perception-wise. Here, we will perform a corpus-based evaluation on a larger corpus to validate our model and to see whether our rules also apply to a more natural conversational setting.

2. Related work

A number of works have addressed the task of automatic BC prediction. Here, we summarize some of the works most relevant to our own research. Based on the observation that backchannels often occur in a pause in Japanese conversations, Noguchi and Den [8] extracted prosodic features in a region of 500ms before the end of pause-bounded phrases. They found that a rise or rise-fall intonation at the end of pause-bounded phrases tend to be followed by backchannels. Cathcart et al. [10] developed a shallow backchannel prediction model based on pause duration and the Part-Of-Speech (POS) tags of the preceding words. Morency et al. [9] used an extensive set of prosodic and lexical features, and eye gaze in combination with advanced sequential probabilistic models. When using machine learning for the prediction of BC timing, it is often difficult to interpret the decision rules which makes generalization to other or broader contexts difficult. Ward and Tsukahara's [4] model detects low regions of pitch and is based on relatively simple rules that provide a great deal of transparency which makes the model easy to understand. In developing our BC prediction model, we strive for the same transparency and simplicity, with the goal to use this BC prediction model in an online setting.

3. Data: IFA Dialog Video Corpus (IFADV)

The IFADV Corpus [12] is a freely available Dutch corpus, containing friendly, spontaneous face-to-face dialogues between well acquainted participants (34 speakers). The participants were allowed to talk about any topic they wanted. Twenty dialogues, totaling 5 hours of speech, were transcribed and annotated on several levels. Additional annotations were made of gaze direction, and functions of dialogue utterances. The latter included annotations of 'minimal response': a minimal response was described in the annotation manual as *'... short responses such as 'yeah' and 'hm' that do not contribute contentful information to the conversation, but are used to maintain the flow of the conversation and to give the interlocutor feedback. Comparable to head nods*. We used these annotations as ground truth labels for BC feedback. Note that although this is an audiovisual corpus, only vocalized minimal responses were annotated; hence, no head nods were annotated. Table 1 shows the distribution of all vocalized minimal responses in the IFADV corpus. We can observe that 'ja' ('yeah') is the most frequent minimal response in

this corpus. Laughter is also often annotated as a minimal response (but not all laughter was annotated as minimal response, the total number of laughters in the corpus is 671). The category 'other' contains various forms of minimal responses; many of these are lexical items or evaluative expressions such as 'precies' ('exactly'), 'juist/klopt' ('right'), 'leuk' ('nice'), etc.

Utterance	Freq.	Utterance	Freq.
ja (yeah)	1498	oh	109
ggg (laughter)	297	oke (okay)	89
hum hum (mm-hmm)	161	ah	13
nee (no)	136	mm	10
other	112	**Total**	**2425**

Table 1: *Vocalized realizations of BC feedback in the IFADV Corpus (English translation in brackets)*

The averaged time interval between 2 BCs produced by a speaker was 14.4s with a standard deviation of 22.0. Furthermore, the word-level transcriptions (automatically aligned) that were available allowed us to look at the contexts in which BCs occur. We were specifically interested in pauses. Of the total of 2425 BCs, 1368 (56%) were placed in a pause. This pause had an average duration of 706ms (standard deviation of 162ms). The remaining BCs overlapped with speech from the interlocutor; 330 (14%) of these overlapping BCs were placed on the last word of the transcribed turn. So, it is fair to conclude that in this corpus, people backchannel in the near proximity of turn-endings: they often backchannel during pauses and often place BCs in anticipation of a turn-ending (as has also been observed in e.g. [4]).

4. Features and Method

We implemented a BC prediction model based on pitch and pause information. We also implemented the well-known BC prediction model by Ward and Tsukahara [4] which will be used for comparison. The models are rule-based to provide transparency and are incremental of nature.

4.1. Pitch & pause model

The pitch and pause model (hereafter referred to as PITCH&PAUSE) is based on the observation that BC feedback often occurs in a pause after a speaker utterance [3], and/or after a falling or rising pitch [3, 7, 5]. In a similar fashion as Ward and Tsukahara [4], we introduce several rules including four parameters: **pauselength**, **pitchslope** (drop and rise), and **lengthpitchslope**. In short, BC feedback is provided upon detection of a pause, and a falling or rising pitch slope, that is preceded by speech. The model is shown in Algorithm 1, with the default values in brackets, and in Fig. 1.

The **pitchsloperise** and **pitchslopedrop** were computed as the difference of pitch between two points of which the length was determined by **lengthpitchslope**. For the detection of pause and speech, we had to come up with practical definitions. We considered a fragment as speech when the fraction of voiced frames was larger than 50% or when the mean intensity of this fragment was above a certain silence threshold. This silence threshold was determined by subtracting 15dB from the mean intensity that was calculated over the whole audiofile. Pause was defined in a similar way: we consider a fragment as a pause when the mean intensity of that fragment is below the previously determined silence threshold or when the fraction of voiced frames is below 15%. The model was implemented by means of a Praat script [13].

Algorithm 1: The PITCH&PAUSE BC prediction model

Provide BC feedback upon detection of:
P1 a pause of **pauselength** (400ms),
P2 preceded by at least 1000ms of speech,
P3 where the last **lengthpitchslope** (100ms),
P4 contain a rising pitch of at least **pitchsloperise** (30Hz) or a falling pitch of at least **pitchslopedrop** (-30Hz).
P5 provided that no BC has been output within the preceding 1400ms.

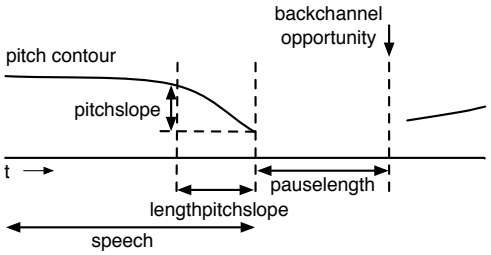


Figure 1: *The* PITCH&PAUSE *model*

4.2. Ward and Tsukahara model

The algorithm introduced by Ward and Tsukahara [4] (hereafter referred to as WARD&TSUKAHARA) is a well-known rule-based algorithm that predicts backchannels based on prosodic features. The model consists of five rules (reprinted from [4]) for the English language:

Algorithm 2: The WARD&TSUKAHARA BC prediction model

Upon detection of
P1 a region of pitch less than the 26th-percentile pitch level
P2 continuing for at least 110ms,
P3 coming after at least 700ms of speech,
P4 providing you have not output backchannel feedback withing the preceding 800ms,
P5 after 700ms wait,
you should produce backchannel feedback.

We implemented this model by means of a Praat script [13]. For speech, we adhered to the same definition as implemented for the PITCH&PAUSE model.

5. Experimental setup

5.1. Data selection

For the estimation of the parameter values of the BC prediction algorithms, we selected 11 fragments from 11 different speakers, each with a duration of 120s, from the left audiochannel (downsampled to 16kHz). These fragments were selected based

on the number of BCs in a time interval of 120s, attempting to reflect the distribution of BCs in the whole corpus, and attempting to obtain an equal number of BCs in each fragment. However, this was not always possible. In the development dataset obtained (hereafter referred to as DEV), the averaged time interval between 2 subsequent BCs is 9.8s (sd of 11.3s). In total, DEV contains 105 annotated BCs. The (disjoint) test set (TEST) was selected in a similar way and is comprised of 8 fragments, each with a duration of 120s, totalling a number of 97 annotated BCs.

5.2. Parameter estimation

In a previous experiment [11], we did not exhaustively estimate the different parameter values in the backchannel prediction models; the main reason was that the main goal of the experiment was to evaluate human perception rather than the models' performances. In the current study, we are interested in comparing the performances between the two models; hence, we search for an optimized set of parameter values. All possible combinations of parameter values were tested and optimized for the F1-measure (see below), averaged over the whole dataset. In addition, we hope to be able to assign meaning to the parameter values found.

5.2.1. PITCH&PAUSE

We tested parameter values in a limited pre-defined range that we considered meaningful. For **pauselength**, we experimented with values of [0.1s, 0.2s, 0.3, 0.4s 0.5s]. For **lengthpitchslope**, the varying values were [0.05s, 0.1s, 0.15s]. Finally, the **pitchslope** was varied between values of [15, 20, 30, 40] for the **pitchsloperise** and **pitchslopedrop**. The default model as was used in [11] will be referred to as PITCH&PAUSE$_{\mathrm{def}}$ and the optimized version as PITCH&PAUSE$_{\mathrm{opt}}$.

5.2.2. WARD&TSUKAHARA

For WARD&TSUKAHARA we experimented with all the parameters (see Algoritm 2). The parameter values were varied as follows: **P1** between [0.24, 0.26, 0.28], **P2** between [0.08, 0.11, 0.14, 0.17], **P3** between [0.3, 0.7, 1.0], **P4** between [0.5, 0.8, 1.1], and **P5** between [0.4, 0.7, 1]. The WARD&TSUKAHARA model as displayed in Algorithm 2 is the 'default' one that has been tested on the English language in Ward and Tsukahara [4]. We will refer to this model as WARD&TSUKAHARA$_{\mathrm{def}}$ and we refer to the optimized version as WARD&TSUKAHARA$_{\mathrm{opt}}$.

5.3. Task and evaluation

One of the problems in evaluating BC prediction models is that BCs are (to a certain extent) optional. A possible solution could be to let multiple participants rate the same fragment on BC opportunities [8, 14]. We are aware of this problem for evaluation but we will not address it in this paper. Here, we will adhere to a 'standard' corpus-based evaluation method, i.e. the predicted output is compared against the reference annotations of a corpus as provided. As evaluation measures, we use precision, recall, and F1-measure. Here, precision is calculated as the number of correct system output BCs divided by the number of all system output BCs. Recall is defined as the number of correctly detected reference BCs divided by the number of all reference BCs. A system output BC is considered correct when the output falls within a certain margin M of the onset of the reference BC (on both sides of the onset). Similarly, a reference BC is correctly detected if there is a system output BC that falls within a certain margin M of the reference BC. The F1-measure is calculated as the weighted mean of precision and recall: $F_1 = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$

6. Results

We report results of the two BC prediction models discussed in this paper: PITCH&PAUSE and WARD&TSUKAHARA. In addition, we look at how different parameter values influence performance.

6.1. Algorithm performance

In Table 2, the performances of the BC prediction models under investigation are presented. In general, the absolute performance is low. However, it is difficult to interpret the absolute performances of the BC prediction models due to the fact that the placement of BCs is optional: some people take the BC opportunity and some do not. In terms of the errors made, this means that a miss or a false alarm may not always be a 'true miss' or a 'true false alarm' which makes interpretation of the performance measure difficult. Therefore, we concentrate on interpreting the performance relative to each other. First, we can observe that there are large differences between the prediction behaviour of PITCH&PAUSE and WARD&TSUKAHARA: PITCH&PAUSE's precision is much higher than its recall, for WARD&TSUKAHARA it is the other way around. The high precision of PITCH&PAUSE is also reflected in the low number of predicted BCs in comparison with WARD&TSUKAHARA. WARD&TSUKAHARA places many unnecessary BCs while PITCH&PAUSE is much more conservative and places BCs at timed moments. This is also visible in Fig. 2 in which F1, precision and recall are shown as a function of the evaluation margin M: the performance of PITCH&PAUSE remains relatively stable while there is a strong increase in performance for WARD&TSUKHARA visible when the evaluation margin is increased.

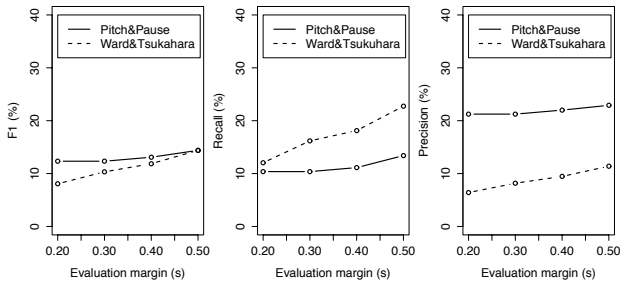
	DEV (N_{BC}=105)			TEST (N_{BC}=97)		
Model	N_{pred}	R	P	N_{pred}	R	P
P&P$_{\mathrm{def}}$	45	2.4	6.2	27	0.1	3.1
W&T$_{\mathrm{def}}$	135	4.6	2.9	106	2.8	4.0
P&P$_{\mathrm{opt}}$	70	10.4	21.2	34	3.5	9.4
W&T$_{\mathrm{opt}}$	180	12.0	6.4	145	8.1	6.6
COMBO$_{\mathrm{opt}}$	227	19.7	9.4	169	9.7	6.9

Table 2: *Results of BC prediction performed on IFADV Corpus, evaluated with an evaluation margin of 0.2s (P=%precision, R=%recall, N_{pred}=the number of BCs predicted, N_{BC} =total number of annotated BCs in the dataset).*

We also combined the output of PITCH&PAUSE and WARD&TSUKAHARA and removed subsequent BCs with an interval $<$ 1s; this slightly increased performance. We further note that optimization of the parameters is needed to achieve a reasonable performance with the risk of loosing genericity of the model.

6.2. Parameter evaluation

The influence of each parameter, when its value is varied while the other parameter values are kept fixed, can be seen in Fig. 3. We can observe that **lengthpitchslope** and **pauselength** are of relatively great influence on performance. The **pauselength**



(a) F1-measure (b) Recall (c) Precision

Figure 2: *Performance on* DEV *expressed as a function of a varying evaluation margin* M.

should not be too short or too long, and the **pitchslope**, which is not too steep, should be measured over a period longer than 100ms.

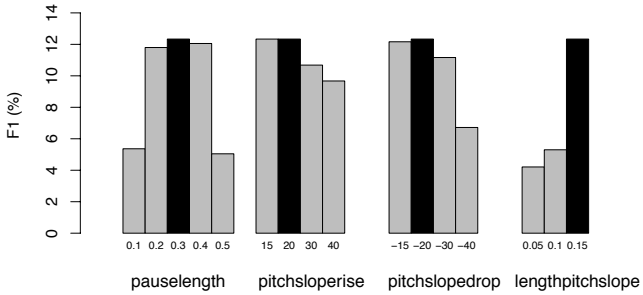


Figure 3: *Influence of* PITCH&PAUSE*'s parameters: the black bars represent the values included in the best performing parameter value set, i.e.* PITCH&PAUSE$_{\text{opt}}$*, with* $M = 0.2$.

For WARD&TSUKAHARA, the most important parameters seemed to be **P1**, **P2**, and **P5** which is in line with what Ward and Tsukahara found [4].

7. Conclusions and Discussion

We have developed a rule-based BC prediction model (in a similar fashion as in [4]) that uses relatively easy extractable and intuitive features such as pitch and pause, and validated it against the Dutch IFADV Corpus of spontaneous free-talk dialogue speech. We have shown that pitch and pause information contribute to the prediction of BCs. Although our PITCH&PAUSE results are slightly better than that of WARD&TSUKAHARA, it is clear that BC prediction is a challenging task. One of the reasons that may have caused this relatively low performance score is the fact that, although the IFADV Corpus is an audiovisual corpus, visual minimal response behaviour was not annotated. Furthermore, corpus-based evaluation of BC prediction is complicated by the fact that BCs are (to a certain extent) optional. Some works [8, 14] have addressed this issue which is a topic for future research. In addition to validating the BC prediction models against a corpus, it is also useful to validate the predicted BCs in a user perception experiment as was done in [11], where it was shown that the PITCH&PAUSE model generated BCs that were perceived nearly as natural as human-generated BCs (see [11] for more details). For future work, we suggest to look more into the observation that BCs are optional: can we identify (prosodic) contexts in which BCs are less 'optional' than other contexts? How does pausing play a role in the gradation of optionality (since we observe that BCs are often placed in pauses and our PITCH&PAUSE model indeed finds this to be an important feature)? Furthermore, what is the width of the time interval in which a BC is still perceived as natural (here, we used a time interval of M=0.2s). Finally, we suggest to look at other easily extractable multimodal speaker cues such as eye gaze for BC prediction.

8. Acknowledgements

This research has been supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet).

9. References

[1] V. H. Yngve, "On getting a word in edgewise," in *Papers from the Sixth Regional Meeting of Chicago Linguistic Society*. Chicago Linguistic Society, 1970, pp. 567–577.

[2] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *Journal of Personality and Social Psychology*, vol. 23, no. 2, pp. 283–292, 1972.

[3] A. T. Dittmann and L. G. Llewellyn, "The phonemic clause as a unit of speech decoding," *Journal of Personality and Social Psychology*, vol. 6, no. 3, pp. 341–349, 1967.

[4] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[5] R. Bertrand, G. Ferré, P. Blache, R. Espesser, and S. Rauzy, "Backchannels revisited from a multimodal perspective," in *Proceedings of Auditory-visual Speech Processing*, 2007, pp. 1–5.

[6] S. Benus, A. Gravano, and J. Hirschberg, "The Prosody of Backchannels in American English," in *Proceedings of the 16th International Congress of Phonetic Sciences 2007*, no. August, 2007, pp. 1065–1068.

[7] A. Gravano and J. Hirschberg, "Backchannel-inviting cues in task-oriented dialogue," in *Proceedings of Interspeech*, 2009, pp. 1019–1022.

[8] H. Noguchi and Y. Den, "Prosody-based detection of the context of backchannel responses," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 1998, pp. 487–490.

[9] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 80–84, 2010.

[10] N. Cathcart, J. Carletta, and E. Klein, "A shallow model of backchannel continuers in spoken dialogue," in *Proceedings of the Conference of the European chapter of the Association for Computational Linguistics - Volume 1*, 2003, pp. 51–58.

[11] R. Poppe, K. P. Truong, D. Reidsma, and D. Heylen, "Backchannel Strategies for Artificial Listeners," in *Proceedings of the International Conference on Interactive Virtual Agents (IVA)*, 2010, p. to appear.

[12] R. J. J. H. van Son, W. Wesseling, E. Sanders, and H. van den Heuvel, "The IFADV corpus: a free dialog video corpus," in *Proceedings of LREC*, 2008.

[13] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer (Version 5.1.07)," 2009. [Online]. Available: http://www.praat.org

[14] L. Huang, L.-P. Morency, and J. Gratch, "Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior," in *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2010, pp. 1265–1272.