A multimodal analysis of vocal and visual backchannels in spontaneous dialogs

Khiet P. Truong, Ronald Poppe, Iwan de Kok, and Dirk Heylen

Human Media Interaction
University of Twente, The Netherlands
{k.p.truong,r.w.poppe,i.a.dekok,d.k.j.heylen}@utwente.nl

Abstract

Backchannels (BCs) are short vocal and visual listener responses that signal attention, interest, and understanding to the speaker. Previous studies have investigated BC prediction in telephone-style dialogs from prosodic cues. In contrast, we consider spontaneous face-to-face dialogs. The additional visual modality allows speaker and listener to monitor each other's attention continuously, and we hypothesize that this affects the BC-inviting cues. In this study, we investigate how gaze, in addition to prosody, can cue BCs. Moreover, we focus on the type of BC performed, with the aim to find out whether vocal and visual BCs are invited by similar cues. In contrast to telephone-style dialogs, we do not find rising/falling pitch to be a BC-inviting cue. However, in a face-to-face setting, gaze appears to cue BCs. In addition, we find that mutual gaze occurs significantly more often during visual BCs. Moreover, vocal BCs are more likely to be timed during pauses in the speaker's speech.

Index Terms: listener response, backchannel, continuer, prediction, head nod, vocalization, gaze, pitch

1. Introduction

During conversations, listeners are continuously signaling attention, interest, and understanding to the speaker through backchannels (BCs) [1]. These short responses do not interrupt the discourse and can be vocal or visual. Vocal BCs include short vocalizations such as 'hmm' or 'uh-huh' while visual BCs include behaviors such as head nods and facial expressions [2].

Research in the area of BCs has mainly focused on identifying speaker contexts that are likely to cue BCs from the listener, and on developing computational models that predict appropriate BCs timings. Observed BC-inviting cues include regions of low pitch, rising or falling final intonations [3, 4], specific part-of-speech tags [5], pauses [5], and energy [4, 6].

Most of these studies have been carried out on telephone-style conversations. In contrast, we consider *face-to-face* conversations where conversants can see each other. The additional visual channel allows listeners to *continuously* express attention, interest and understanding without interfering with the vocal channel. It is therefore likely that BC behavior differs between the telephone-style and face-to-face setting. Specifically, we expect that BC-inviting cues are different as they might be expressed over the visual channel instead of or in combination with the vocal channel. In addition, there might be a relation between BC-inviting cue and the modality of the BC produced, i.e. vocal or visual.

Some researchers have addressed differences between vocal and visual BCs [7, 8], but not in relation to BC-inviting cues. Morency et al. (2010) [9] used gaze in addition to speech features for the automatic prediction of BCs but did not consider the type of BCs. Bertrand et al. (2007) [10] considered speaker gaze as a BC-inviting cue in relation to the BC modality, but a quantitative analysis of frequency and timing was not part of their study.

Our long-term goal is to provide Embodied Conversational Agents (ECAs) with the ability to display human-like listening behavior. This requires, on the one hand, automatic prediction of BC timing in response to detected BC-inviting cues. On the other hand, given the audiovisual nature of the ECA, we need to generate the BC in the proper modality.

In this paper, we take a step towards this goal by investigating the relation between BC-inviting cues and the modalities of the BCs using corpus analysis. Given our corpus of face-to-face dialogs, we analyze the prosodic and gazing behavior of speakers and listeners in the vicinity of vocal and visual BCs, with the aim to find out whether vocal and visual BCs are invited by similar cues. Since our eventual goal is to predict BCs online, we investigate cues that are easily automatically detectable. With regard to vocal cues, we look at pause, energy, and pitch features from the speaker. With regard to visual cues, we consider gaze behavior from both the speaker and the listener. As BCs can have several functions, both communicative and affective, we restrict ourselves to the function as *continuer* signal indicating that the speaker should go on talking, and discard the more complex types of BCs such as repetitions or completion of sentences. We only consider head nods and shakes as visual BCs as other visual BCs such as smiles or frowns usually carry an additional attitudinal or affective function.

The paper is structured as follows. Section 2 describes the audiovisual corpus and annotations. Corpus analysis and results are discussed in Section 3 . We conclude in Section 4.

2. IFADV corpus

2.1. Recordings

The IFADV corpus [11] is a publicly available audiovisual Dutch corpus, containing spontaneous face-to-face dialogs between acquainted participants. The dialogs are unscripted and no task was given; the participants were allowed to talk about any topic. Each dialog has a duration of 15 minutes. Of the 20 annotated dialogs available, one dialog was discarded because the conversants were constantly aware of being recorded.

2.2. Annotations of backchannels

For vocal BCs, we used the conversational function annotations provided with the corpus. These annotations contain the class 'minimal response' which were used to locate all vocal BCs. However, because the definition of minimal response in the IFADV annotation manual encompasses more than only the continuer function, these minimal response annotations were in-

spected and adjusted accordingly. In particular, we removed laughter sounds that were annotated as minimal responses.

Annotations of visual BCs were *not* provided with the IFADV corpus. Therefore, the dialogs were annotated on head nods and shakes with a clear continuer function. The *start* of the head movement was annotated by marking the beginning of the stroke, the most effortful part of the head movement. Repeated nods were annotated as separate nods when there was a visibly large change in amplitude or velocity in the nod movement. First, three annotators coded the same dialog and discussed the instances that were disagreed upon. After clearing up the disagreements, each of the annotators coded 6 or 7 dialogs. The annotators checked each other's annotations and adjusted these when appropriate.

2.3. Frequency of backchannels

In total, we identified 3283 BCs in 19 dialogs of the corpus. We make a distinction between BCs that are produced by either a vocal or a visual expression and BCs that are produced both vocally and visually simultaneously. We refer to this latter group as bimodal BCs. A BC was considered bimodal when a vocal and a visual BC both occurred within a margin of 0.2 s (to the right and the left). This criterion results in 430 (13%) bimodal BCs, 1596 (49%) vocal BCs, and 1257 (38%) visual BCs.

In our analyses, we only used the start times of the BCs. Despite previous observations that a nod precedes a vocal BC by 175 ms on average [7], we found no systematic difference between the vocal and visual onset of bimodal BCs. We used the earliest onset of the two.

3. Corpus analysis

In our analysis, we look at cues that have been cited frequently as BC-inviting and can be automatically extracted: prosody and gaze. Pause, energy, and pitch features were extracted from the interlocutor's speech, while gaze was extracted from both the interlocutor and the person producing a BC (which we refer to as BC-er).

3.1. Speech activity and pause

BCs have been found to occur near the ending of clauses [3] and are frequently placed during the interlocutor's speech. In practice, we also observe that BCs are frequently placed in within-turn or between-turn pauses of the interlocutor [5, 12]. Since head nods are not interfering with the interlocutor's speech, one would expect that they are placed throughout the discourse whereas vocal BCs more often occur in pauses. Hence, we look at the placement of BCs in speech or pause. The manually chunk-aligned orthographic annotations from the IFADV corpus, sampled at 0.01 s, were used as ground-truth speech activity labels. We also automatically determined speech/non-speech segmentations with similar results. However, in our analyses we used the manual annotations as these are provided with the corpus.

First, we investigate the placement of BCs within the interlocutor's speech segment, to see whether visual BCs are more uniformly distributed than vocal BCs throughout a speech segment. From Fig. 1, it is clear that this is not the case. These distributions vary slightly for different speech segment lengths. Vocal and visual BCs have similar distributions with increasing probability as the speech progresses. We expect that the notion of attention becomes more important towards the end of the speech turn.

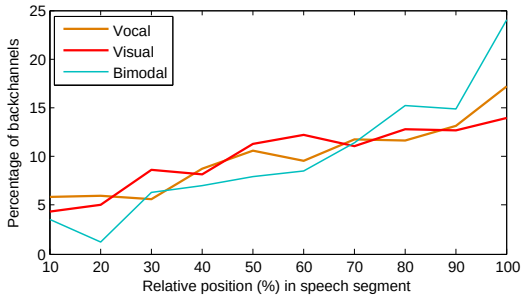

Figure 1: *BCs relative within the interlocutor's speech segment*

Next, we look at the percentages of BCs that fall within a pause. For vocal and visual BCs this amounts to 37% and 16%. These numbers are significantly higher and lower than the chance of 28.4 % that a BC falls in a pause ($\chi^2(1) = 60.32, p < 0.001$ and $\chi^2(1) = 89.06, p < 0.001$, respectively). We take a closer look at the vicinity of BCs with respect to the average amount of interlocutor's speech.

In Fig. 2, we observe differences between vocal and visual BCs: vocal BCs are surrounded by a lower percentage of speech frames than visual BCs. This difference is largest around the start of a BC. Also it can be observed that the average amount of speech is lower after a BC which indicates that BCs are often placed slightly before or after the end of a speech segment.

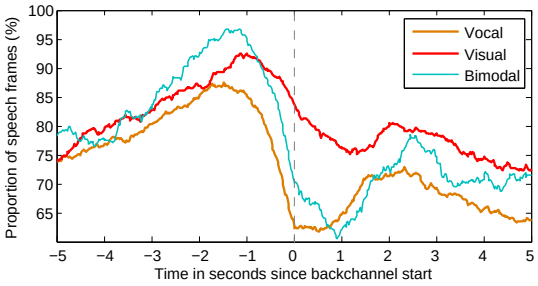

Figure 2: *Amount of interlocutor speech in the vicinity of a BC*

3.2. Pitch

Pitch was extracted automatically using Praat [13] and sampled at 0.01 s. Following [3], intervening speech regions where no pitch was detected (up to 80 ms) were linearly interpolated to spread pitch values over consonant sounds and to correct for frames missed by the pitch extractor. When at least 50% of all frames in a time period contained pitch values, the contour was considered valid and a least-squares linear regression model was fitted. There is no pitch during pauses and as vocal BCs are more often timed in the interlocutor's pause, the number of valid pitch contours preceding vocal BCs is lower (35.9% versus 63.4% before visual BCs for a region of 0.11 s).

Following literature on prosodic BC-inviting cues, we considered the low pitch cue and rising and falling pitch contours. A low pitch region was defined by Ward and Tsukahara (2000) [3] as a period of 110 ms (we also inspected 300 ms) in which all pitch frames are below the 26^{th} percentile of the speaker's pitch values. In Table 1, we observe that a low pitch

region indeed occurs significantly more often than chance before a BC ($\chi^2(1) = 74.91, p < 0.001$). However, no large differences between vocal and visual BCs were observed. Additionally, we analyzed regions of high pitch (pitch levels above the 74^{th} percentile) but found that BCs are not likely to be invited by this cue.

	110 ms		300 ms	
	low	high	low	high
Chance	18.1	17.7	13.2	12.2
Any BC	26.5	14.6	17.3	10.4
Vocal	28.1	16.6	19.9	13.1
Visual	25.9	13.9	15.5	8.8
Bimodal	24.4	11.5	16.6	8.9

Table 1: *BCs preceded by a region of low/high pitch (%)*

Rising and falling pitch contours are commonly found to invite BCs [3, 4]. For valid pitch contours, we subtracted the last pitch value from the first pitch value. Differences above 30 Hz or below -30 Hz were considered rising and falling pitch slopes, respectively. In contrast to previous studies on telephone-style dialogs, we observe in Table 2 that rising or falling pitch are not strong BC-inviting cues; the probability that these cues occur before a BC is not or hardly above chance. Furthermore, no differences between vocal and visual BCs are found. We attribute this discrepancy to the additional visual channel in face-to-face conversation. As the listener can use (mutual) gaze to convey attention, the need to cue BCs over the vocal channel is reduced.

	110 ms		300 ms	
	rising	falling	rising	falling
Chance	12.9	9.5	24.0	18.0
Any BC	11.7	9.6	20.6	21.3
Vocal	11.9	9.6	19.6	22.8
Visual	10.7	9.4	20.9	19.7
Bimodal	14.8	10.6	22.6	23.4

Table 2: *BCs preceded by a region of rising/falling pitch (%)*

3.3. Energy

High energy and falling energy slopes in the interlocutor's speech have been found to precede BCs [4, 6]. We treat energy similarly to pitch by fitting a least-squares linear regression model to a 0.3 s window. High energy profiles are windows with energy levels above the 74^{th} percentile, and are usually perceived as emphasized utterances. We found that the number of visual BCs preceded by a high energy profile was significantly higher than vocal BCs (18.1% and 9.7%, respectively, $\chi(1) = 183.94, p < 0.001$), with chance being 11.0%. These numbers might indicate that visual BCs are, more often than vocal BCs, cued by emphasized utterances in the interlocutor's speech. In addition, 24.9% of the visual and 33.1% of the vocal BCs is preceded by an energy decrease of at least 10 dB at a chance level of 21.2%. These findings strongly correlate with the more frequent occurrence of vocal BCs in pauses of the interlocutor's speech, and the observation that a speech segment usually ends with a falling energy slope.

3.4. Gaze

In face-to-face conversations, gaze is an important cue that regulates turn-taking and elicits attention signals [14]. We expect a relation between gaze and the modality of the BC produced. Specifically, we expect that the number of visual BCs during mutual gaze will be higher as the BC-er knows that the interlocutor will see his signal.

In the analysis, we used the gaze annotations provided with the IFADV corpus, sampled at 0.01 s. These annotations provide, at each time instant and for each subject, whether or not there is gaze at the other. When both subjects gaze at the other, there is mutual gaze. While speaking, the interlocutor gazes in 68.3% of the time at the BC-er. Conversely, the interlocutor is being looked at by the BC-er in 85.1% of the time. As we assume that the BC-er is not speaking, we use these numbers in Table 3 as the chance levels for when the interlocutor is gazing, or being looked at, respectively. This asymmetry in the amount of gaze from and towards the interlocutor is frequently reported in literature (e.g. [14]).

	gaze at BC-er	gaze from BC-er	mutual gaze
Chance	85.1	68.3	57.3
Any BC	86.7	85.0	75.2
Vocal	82.0	81.0	68.4
Visual	92.2	88.2	82.0
Bimodal	87.9	90.5	80.5

Table 3: *BCs during gaze (%)*

In 57.3% of the time, there is mutual gaze. Fig. 3 shows the average amount of mutual gaze in the vicinity of a BC. As witnessed by the initial increase in mutual gaze before a BC, mutual gaze appears to be a strong cue in eliciting BCs. We see similar gaze patterns for the interlocutor and BC-er individually. The gaze cue seems stronger for visual and bimodal BCs compared to vocal BCs, as we hypothesized. Slightly before and during the BC, the amount of mutual gaze decreases. However, we still expect that the (re-)occurrence of gaze, rather than the aversion of gaze, is the cue for the BC, and that there is some response time before the BC is produced.

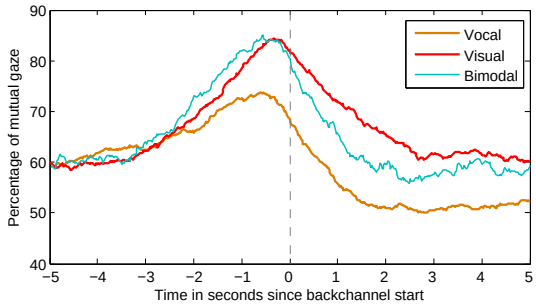

Figure 3: *Amount of mutual gaze in the vicinity of a BC*

After a vocal BC, the average amount of mutual gaze is systematically lower compared to visual BCs. We observe that the average amount of mutual gaze during pauses is 46.8%, in comparison to 59.1% mutual gaze during speech. Given the higher number of vocal BCs in pause, this might explain the lower amount of mutual gaze after the vocal BC. An alterna-

tive explanation comes from observations made on turn-taking behavior. Backchannels can be regarded as a response by the listener without the intent of taking the turn. In this respect, the interlocutor's gaze aversion might be seen as an anticipation of a potential but unwanted claim of the turn by the BC-er. A more careful analysis of differences in context between BCs and turn claims, such as Koiso et al. (1998) [6] performed for telephone-style dialogs, in relation to the type and timing of behaviors is necessary.

4. Conclusions

We have analyzed an audiovisual corpus of spontaneous face-to-face dialogs to investigate the relation between subjects' vocal behavior and gaze on the timing of backchannels (BCs). Specifically, we analyzed whether visual and vocal BCs were preceded by different cues.

We found notable differences between the face-to-face dialogs that we considered, and previous findings from telephone-style dialogs. A rising or falling pitch before a BC, reported to be a strong BC-inviting cue [3, 4, 10], did not occur more often than chance. We expect that the additional visual modality affects BC behavior. In particular, (mutual) gaze and facial expressions can be used to signal attention to the interlocutor. This reduces the need to vocally express attention as in a telephone-style setting. Indeed, slightly before a BC, the amount of gaze increases, which makes it a good predictor for BCs.

We also observed differences in the relation between BC-inviting cues and the type of BC, i.e. visual or vocal. Firstly, visual BCs are more likely to be timed during the interlocutor's speech than vocal BCs. Secondly, we observed that they were more often preceded by a high energy profile in the interlocutor's speech. It might be that emphasized speech utterances cue visual BCs, but this requires further investigation. As speech energy drops before pauses, and given the more frequent occurrence of vocal BCs in pauses, it is clear that decreasing energy is a good cue for vocal BCs. Another difference between the occurrences of vocal and visual BCs was found for (mutual) gaze. Mutual gaze appears to be a stronger cue for visual BCs, compared to vocal BCs. Probably, this is because the BC-er knows that his signal will be seen by the interlocutor. During and after the BC, the amount of gaze is lower. After vocal BCs, this effect is even more prevalent.

In summary, the form of BCs, visual or vocal, depends to some extent on the context in which they occur. We also considered the class of bimodal BCs, where a vocal and visual BC are produced with almost equal timing. Bimodal BCs appeared to have intermediate characteristics between vocal and visual BCs. Our goal is to develop Embodied Conversational Agents (ECAs) that can listen attentively to a speaker and generate appropriate feedback signals. To this end, we can formulate 'rules' to predict not only the timing of BCs, but also their type (i.e. visual or vocal). In accordance with our results, we can, for example, decide to display a visual BC when this BC is predicted during the interlocutor's speech.

Given some notable differences with telephone-style dialogs considered in previous studies, we plan to further analyze vocal and visual BC behavior in spontaneous face-to-face dialogs. Specifically, we will consider visual BCs with an attitudinal or affective function that were discarded in the current study, e.g. frowns and smiles. Because of their potentially different roles, they might be displayed in different contexts.

In the research described in this paper, we have used both manual and automatic annotations, with in mind that the features used can be extracted automatically. However, online processing of the interlocutor's speech and visual behaviors will continue to be a challenging task, especially given the limited processing time available between the observation of a cue, and the BC that might follow. Eventually, we target automatic online prediction of BCs, using either a rule-based (e.g. [3, 12]) or machine learning-based (e.g. [9]) approach.

Finally, we plan to investigate the interchangeability of BC types. For example, in certain cases, a vocally performed BC might also have been a nod, and *vice versa*. This issue cannot be answered using corpus research, and we intend to use human perception studies along the lines of [15] instead.

5. Acknowledgements

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

6. 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 Jr., "On the structure of speaker-auditor interaction during speaking turns," *Language in Society*, vol. 3, no. 2, pp. 161–180, 1974.

[3] 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.

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

[5] N. Cathcart, J. Carletta, and E. Klein, "A shallow model of backchannel continuers in spoken dialogue," in *Proceedings of EACL*, 2003, pp. 51–58.

[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 and Speech*, vol. 41, no. 3-4, pp. 295–321, 1998.

[7] A. T. Dittmann and L. G. Llewellyn, "Relationship between vocalizations and head nods as listener responses," *Journal of Personality and Social Psychology*, vol. 9, no. 1, pp. 79–84, 1968.

[8] J. Allwood and L. Cerrato, "A study of gestural feedback expressions," in *Proceedings of the First Nordic Symposium on Multimodal Communication*, 2003, pp. 7–22.

[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, pp. 70–84, 2010.

[10] R. Bertrand, G. Ferré, P. Blache, R. Espesser, and S. Rauzy, "Backchannels revisited from a multimodal perspective," in *Proceedings of AVSP*, 2007.

[11] 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, pp. 501–508.

[12] K. P. Truong, R. Poppe, and D. Heylen, "A rule-based backchannel prediction model using pitch and pause information," in *Proceedings of Interspeech*, 2010, pp. 3058–3061.

[13] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[14] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[15] R. Poppe, K. P. Truong, and D. K. J. Heylen, "Backchannels: quantity, type and timing matters," in *Proceedings of IVA*, 2011, to appear.