Backchannel Strategies for Artificial Listeners

Ronald Poppe, Khiet P. Truong, Dennis Reidsma, and Dirk Heylen*

Human Media Interaction Group, University of Twente
P.O. Box 217, 7500 AE, Enschede, The Netherlands
{poppe,truongkp,dennisr,heylen}@ewi.utwente.nl

Abstract. We evaluate multimodal rule-based strategies for backchannel (BC) generation in face-to-face conversations. Such strategies can be used by artificial listeners to determine when to produce a BC in dialogs with human speakers. In this research, we consider features from the speaker's speech and gaze. We used six rule-based strategies to determine the placement of BCs. The BCs were performed by an intelligent virtual agent using nods and vocalizations. In a user perception experiment, participants were shown video fragments of a human speaker together with an artificial listener who produced BC behavior according to one of the strategies. Participants were asked to rate how likely they thought the BC behavior had been performed by a human listener. We found that the number, timing and type of BC had a significant effect on how human-like the BC behavior was perceived.

1 Introduction

We introduce and evaluate strategies to automatically generate listener backchannels in face-to-face interactions between a human speaker and an artificial listener. In our case, the artificial listener is an intelligent virtual agent that aims at sustaining a conversation with a human interlocutor. Backchannels (BCs) are an important aspect of conversation management, specifically for this type of speaker-listener dialog. Their function is to signal attention and interest, without interrupting the speaker's discourse. The use of appropriately timed BCs has been found to improve the speaker's narrative, and increase the amount of time spent speaking [1]. In interactions between a human speaker and an artificial listener, BCs should be automatically generated for the artificial listener. While placement of BCs within the discourse structure is reasonably well understood (see e.g. [2,3]), it is difficult to determine and analyze this lexical structure in real-time. In practice, one has to resort to features that can be obtained with low processing requirements such as gaze direction, pitch slopes and pauses in the speaker's speech (e.g. [4,5]).

Several authors have addressed real-time prediction of BC timing in audio-only settings using machine learning [6,7] or rule-based algorithms [8]. Machine

* The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/20072013) under Grant agreement no. 211486 (SEMAINE).

J. Allbeck et al. (Eds.): IVA 2010, LNAI 6356, pp. 146–158, 2010.

learning algorithms can automatically extract decision rules from labeled training samples. These samples must be representative of the target domain. In practice, it is often difficult to interpret the decision rules, which makes generalization to other contexts difficult. This would require retraining and labeled samples must be available for the new domain. To avoid these issues, Ward and Tsukahara [8] manually defined a rule-based algorithm that predicts BCs based on the speaker's pitch contours. They obtained reasonable results while the algorithms are easy to understand and verify.

The above works have considered an audio-only setting, which is different from the face-to-face setting with an artificial listener that we consider in this research. For example, Dittmann and Llewellyn [9] observed that the mere fact that conversational partners can see each other is a signal of attention and thus reduces the need for BCs. Also, the turn-taking process, including BC feedback, is arguably more complex. Duncan [2] identifies speech, gaze and gesture as relevant components of natural turn-taking, while turn-taking in audio-only settings only considers speech. Morency *et al.* [10] presented one of the few real-time multimodal BC prediction systems. They automatically select features from speech and gaze, and train conditional random fields to model sequential probabilities.

Systems that automatically predict BC timings are usually evaluated using the correlation between the predicted and actually performed BCs. This approach does not take into account individual differences. A given moment where an individual does not provide a BC is not necessary an inappropriate moment for BC feedback. However, in the evaluation of BC prediction algorithms, such a predicted BC would be regarded as a false positive. Individual differences also affect the training of machine learning models as samples without BCs are labeled as negatives which consequently decreases the quality of the model. This issue was addressed by Huang *et al.* [11], who had observers watch a video of a speaker and indicate where they would provide BCs as if they were actual listeners. They analyzed which BC opportunities were shared by several raters and used these samples as positives. The output of their approach was rated by human observers on believability, rapport, wrong head nods and missed opportunities. This evaluation based on human perception is in contrast with corpus-based evaluation. Even though one would expect that the actual listener would demonstrate human-like BC behavior, there is no guarantee that the behavior will be judged as most human-like. Indeed, Huang *et al.* found that a face-to-face setting was not best rated by human observers.

In this paper, we focus on face-to-face conversations with an artificial listener and introduce and evaluate several strategies that define BC timings using intuitive rules. We introduce strategies that use features from the speaker's speech and gaze. In addition to multimodal input, we also deliberately generate both visual and vocal BCs. The strategies are evaluated with a human perception experiment where participants judge the naturalness of the artificial listener's BC behavior, performed by a virtual agent. Our samples are short, which allows for a closer analysis of the relative strengths of the different strategies. We investigate the influence of the amount, timing and type of BCs.

We discuss related work on BCs and artificial listeners in the next section. The six BC strategies that we will evaluate are described in Section 3. We present our experiment in Section 4 and discuss the results in Section 5.

2 Related Work

Research into turn-taking behavior defines the person who holds the turn as the speaker and the person who is being addressed as the listener. The term BC feedback was first used by Yngve [3], who described it as messages sent by the listener without the intent to take the turn. Subsequent research focused on the role of BCs within the turn-taking process. Duncan [12] observed that BCs are often used as a response to a speaker's turn-yielding signal, implying that they might be cued by the speaker. BCs take many forms including short vocalizations (e.g. "hmm", "uhhuh"), sentence completions, requests for clarification, brief restatements and bodily manifestations such as head nods [2]. Bavelas *et al.* [1] identified specific and generic responses. The former are tightly connected to the speaker's narrative, the latter are mere signals of continued attention.

Apart from the role of BCs in conversation, researchers have focused on identifying the nonverbal context of BCs. Dittmann and Llewellyn [13], Duncan [2] and Yngve [3] noted that BCs are often produced after rhythmic units in the speaker's speech, and specifically at the end of grammatical clauses. Kendon [14] and Bavelas *et al.* [15] looked at the relation between gaze and BCs, and found that the listener is likely to produce a BC when there is a short period of mutual gaze between speaker and listener. These moments usually occur at the end of a speaker's turn.

Corpus-based research revealed that BCs are also often produced when the speaker's speech ends with a rising or falling pitch [5]. Bertrand *et al.* [4] additionally took into account gaze and lexical features and observed that BCs often appear after nouns, verbs and adverbs. Cathcart *et al.* [16] further found that BCs are often preceded by a short pause in the speaker's discourse.

These systematics have motivated researchers to train machine learning models to automatically and in real-time predict the timing of BCs given the speaker's discourse. For instance, such systems can be used to insert BC feedback in spoken telephone dialogs. Given the real-time requirements, these systems have to rely on audio features that are easy to extract. Noguchi [6] learned decision trees based on pitch and power features. Okato *et al.* [7] used Hidden Markov Models where state transitions corresponded to changes in prosodic context. Instead of using a learned model, Ward and Tsukahara [8] manually defined an algorithm for the prediction of BC timing using a small number of rules. It focuses on a sustained lowered pitch within the speaker's utterance.

In face-to-face conversations between a human speaker and an artificial listener, also visual cues from the speaker should be taken into account. Morency *et al.* [10] automatically selected audio and gaze features extracted from a corpus of training samples. For the prediction of BC timings, they trained conditional random fields, a sequential probabilistic model. Maatman *et al.* [17] aimed at creating rapport between an artificial listener and a human speaker. They observed several aspects of the speaker's nonverbal behavior, including speech pitch and loudness, gaze, head nods and shakes and posture shifts. A modification of Ward and Tsukahara's algorithm [8] was combined with a mimicking strategy to identify BC opportunities. The approach was evaluated in [18] and was found to improve the overall impression of the communication compared to a similar setting without the animation of BCs. However, they only used head nods as BC type. Huang *et al.* [11] also conducted a human perception experiment to evaluate the quality of their BC strategy, but they also only animated nods.

In natural face-to-face conversations, people use a range of different BC types including nods, vocalizations, smiles and frowns [2]. There is some research on how different BC types are perceived in terms of positivity, politeness and attentiveness. For example, Heylen *et al.* [19] investigated the perception of facial expressions as BCs. Granström *et al.* [20] also took into account systematic variations in the pitch contour of vocalizations. However, this research considered isolated BCs without context such as the speaker's discourse. Given differences in perceived meaning of BC types, it would make sense to explicitly take into account BC type in the generation of an artificial listener's BC behavior. In the research described in this paper, we evaluate several rule-based BC strategies using a human perception experiment. Participants rated fragments of a conversation between a human speaker and an artificial listener who displayed BC behavior according to one of the strategies. We used two different BC types (a nod and a vocalization). We investigate how the amount, timing and type of BC influences the perception of naturalness of the BC behavior.

3 Backchannel Strategies

We define six strategies to determine the placement of listener BCs in real-time based on the speaker's speech, gaze or both. We discuss the strategies below.

- **Copy.** This strategy contains all BCs that have been performed by the actual listener in interaction with the speaker.
- **Random.** An approximation of an Erlang distribution is used to generate BC timings without taking into account any signal from the speaker. We use one normal distribution to model the timing of the first BC, and one to model the time between two BCs. For the generation of BC onsets, we iteratively sample from the distributions until the end of the fragment is reached. We resample when the time between subsequent BCs is below 1s. One random distribution for each strategy-fragment combination is generated.
- **Ward & Tsukahara.** We use the rule by Ward and Tsukahara [8] that has been used for BC prediction in English audio-only settings, reprinted as Algorithm 1.

Algorithm 1. WARD&TSUKAHARA strategy for BC placement

```
Provide BC feedback upon detection of:
P1 a region of pitch less than the 26th-percentile pitch level and
P2 continuing for at least 110ms,
P3 coming after at least 700ms of speech,
P4 provided that no BC has been output within the preceding 800ms,
P5 after 700ms wait.
```

- **Gaze.** Several researchers have observed the relation between (mutual) gaze and BCs. In a setting similar to ours, Bavelas *et al.* [15] observed that listeners tend to look at the speaker for fairly long intervals, while speakers would look at the listener for frequent but much shorter periods. When the speaker looks at the listener, this starts a brief period of mutual gaze, in which a BC is likely to occur. Similar observations have been made by Duncan [12] and Kendon [14]. We use this mechanism to determine the timing of BCs based on the speaker's gaze at the listener only, formalized in Algorithm 2.

Algorithm 2. GAZE strategy for BC placement

```
Provide BC feedback upon detection of:
P1 gaze at the listener,
P2 coming after at least 1000ms of no gaze,
P3 after 500ms wait.
```

- **Pitch & Pause.** It has been observed that BCs frequently occur in a pause after a speaker's utterance [3,13]. We include this observation in our strategy for BC placement in Algorithm 3. We use a minimum pause duration of 400ms as a compromise between the 200ms as used in Maatman *et al.* [17] and the 700ms used by Ward and Tsukahara [8]. We further take into account the preceding speech. Instead of a region of low pitch, we focus on rising or falling pitches, as suggested by several researchers [9]. Gravano and Hirschberg [5] found in their audio-only corpus that over 80% of all BCs were preceded by either a rising or falling pitch contour.

Algorithm 3. PITCH&PAUSE strategy for BC placement

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

- **Pitch, Pause & Gaze.** In this strategy, we combine the BCs of the GAZE and PITCH&PAUSE strategies, both described above. Our rationale is that both should identify relevant BC locations. To avoid overlapping BCs, we set the minimum time between two BCs to 1s. In a situation where both strategies identify the same locations, the combined strategy will result in similar placement of BCs.

4 Experiment Setup

We conducted a user experiment where human observers rated fragments from actual conversations. We replaced the listener by an artificial listener and generated BC behavior according to the strategies described in the previous section. In this section, we explain the construction of the stimuli and the setup of the experiment.

4.1 Stimuli

We used the Semaine Solid SAL data [21], which contains dialogs between a human listener and a human speaker. The task of the listener was to sustain the conversation with the user, while playing one of four predefined roles with the intention to evoke emotionally colored reactions. We only consider interactions with Prudence, the pragmatic character, played by two different operators.

We extracted speaker fragments bounded by 1.5s of non-speech. We discarded fragments that were shorter than 10s, did not contain BCs or that contained interjections (e.g. “that’s good”). From the remaining fragments, we selected 8 samples for each of the two operators. We further removed the listener’s vocalizations from the speaker’s audio signal. The average sample length was 19.9s, with an average of 2.6 BC per fragment.

Speaking/pause and pitch information were obtained using Praat [22], speaker’s gaze towards the listener was annotated manually. For the COPY strategy, we annotated the onset of all BCs. In the case of repeated nods, we took the most articulated nod. When two BCs overlapped in time (e.g. nod and vocal), we annotated a single instance. For the RANDOM strategy, the mean (SD) start of the first BC was at 6.97s (6.20s), and 5.00s (2.73s) between subsequent BCs.

In order to evaluate the BC strategies, we replaced the video of the listener by a virtual agent. We used Elckerlyc [23], a BML realizer that allows for easy control of verbal and nonverbal behavior of the artificial listener. Given that we do not focus on the content of the speaker’s utterance, we choose two common generic (see [1]) BCs in face-to-face interaction: a nod and vocalization (“uh-huh”). There is surprisingly little known about potential semantic differences between visual and vocal BCs. Duncan [2] found no difference in the placement within the speaker’s discourse. This was also observed by Dittmann and Llewellyn [9], who did note that a nod on average precedes a vocalization by 175ms. The large number of BCs in our sample sets did not allow for a controlled design and we introduced BC type as an uncontrolled variable in our experiment. At each defined BC onset, we randomly animated a nod, a vocalization or a combination of both. We used the same ratios as performed by the actual listeners, calculated over all fragments.

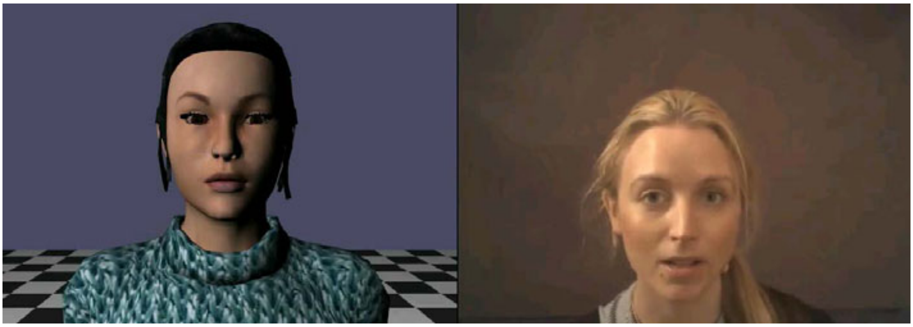

Fig. 1. Example stimulus with artificial listener (left) and actual speaker (right)

We also animated, for each fragment and strategy, the listener's blinks where they occurred in the actual recording. The rationale for this decision is that blinks can sometimes be regarded as BCs, but it is unclear to what extend they can be replaced by a different type of BC. In addition, the use of blinks prevents the artificial listener from looking too static. The final stimuli consisted of the animated listener and the video of the speaker, shown side-by-side (see Figure 1).

4.2 Procedure

The participants were explained they would be participating in an experiment to determine the quality of BC strategies. They were told they would be shown fragments of a conversation between a speaker and an animated listener who would only show nods and blinks, and say "uhhuh". Participants were asked to rate, for each fragment, "how likely do you think the listener's backchannel behavior has been performed by a human listener". They made their judgements by setting a slider that corresponded to a value between 0 and 100. For each fragment, the participants could type in optional remarks. After completing the experiment, they were asked to provide general comments on the study.

Given the large number of combinations (16 fragments and six strategies), we divided fragments into two distinct sets. Each set contained four fragments of each of the two operators, performed with all six BC strategies. Each participant therefore rated 48 samples. We defined a pseudo-random order, with the only constraint that a fragment would not appear twice in succession. Half of the participants viewed the clips in the specified order, the other half in the reverse order. Order and set were crossed to yield four groups.

Due to the different sets of fragments, we formally have two experiments, one for each set. Each experiment has strategy and fragment as within-subjects variable and order as between-subjects variable.

4.3 Participants

We recruited 20 colleagues and doctoral students (4 female, 16 male) with a mean age of 28.4 (min 24, max 55). Each of the participants was assigned randomly to a group with the lowest number of respondents.

5 Results and Discussion

The 20 participants rated in total 960 samples. Initially, we ignored the variable of fragment which allowed us to combine the results of both sets of fragments. We performed a repeated measures ANOVA with set and order as between-subjects variables, and strategy as within-subjects variable. The average score per strategy for each participant was used as the dependent variable. In the analysis, only the variable strategy proved significant (F(5, 80) = 22.141, $p < 0.01$). See Figure 2 and Table 1 for an overview and the scores per strategy. Post-hoc analysis revealed significant differences between all pairs of strategies ($p < 0.05$), except between the RANDOM, PITCH&PAUSE and PITCH,PAUSE&GAZE strategies. The high SDs for each strategy are partly explained by differences in rating scores of individual participants. While the average score over all samples was approximately 42, the range for individual participants was between 17 and 62. An analysis with (normalized) z-scores resulted in the same interaction effects.

When looking at the scores of the different strategies, our first observation is that the COPY strategy performed the best on average. This is not surprising as the timing of BCs was performed by the actual listener, although in many cases the animated BC type was different from the type that was actually performed. The relatively low score of the COPY condition might partly be attributed to the BC type. We used two generic BC types, which might not be the most suitable choice in all cases. Also, we expect that part of the lower score can be attributed to inter-personal differences in BC behavior. Several participants reported that they performed BCs based on the speaker's video as if they were the listener.

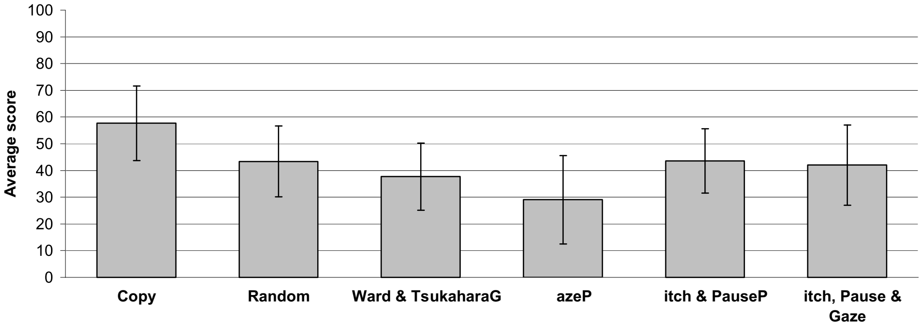

Fig. 2. Average scores (with SD) for all strategies, calculated over all responses

Table 1. Summary of results per strategy. A generated BC is matching if its onset appears within a margin of 200ms of the onset of a BC in the COPY strategy

	COPY	RANDOM	WARD& TSUKAHARA	GAZE	PITCH& PAUSE	PITCH,PAUSE & GAZE
Average score	57.67	43.39	37.69	29.04	43.56	42.03
Standard deviation	13.93	13.26	12.60	16.57	12.02	14.98
Number of BCs	51	47	54	25	25	49
Nods (%)	41.18	48.94	50.00	64.00	48.00	46.94
Nod + vocals (%)	58.82	48.94	50.00	36.00	52.00	53.06
Vocals (%)	0.00	2.13	0.00	0.00	0.00	0.00
Matching precision (%)		12.77	9.26	4.00	40.00	22.49
Matching recall (%)		11.76	9.80	1.96	19.60	21.57

Simultaneously, they monitored the animation of the listener to see how well the animated BCs corresponded with their own BCs.

We further observe that the RANDOM, PITCH&PAUSE and PITCH,PAUSE&GAZE strategies performed relatively well, while the WARD&TSUKAHARA and GAZE strategies received significantly lower scores.

Number of BCs. In a first attempt to explain these differences, we analyzed the effect of the number of BCs. We calculated the correlation between the number of BCs per minute and the average score, for each fragment and each strategy. The effect appeared to be significant ($r(94) = 0.400$, $p < 0.01$). The range of BCs per minute was between 0 and 23. Given this correlation, it is more likely that strategies with a lower number of generated BCs will have lower scores. This is true for the GAZE strategy, but the PITCH&PAUSE strategy scored similar to the RANDOM strategy while only half the number of BCs was generated (see Table 1). Also, the additional BCs of the GAZE strategy did not increase the score of the PITCH,PAUSE&GAZE strategy compared to the PITCH&PAUSE strategy. Clearly, the number of BCs is not the only important factor.

Timing of BCs. To quantitatively investigate the timing of the BCs, we calculated how well these matched the BCs in the COPY strategy. While an unmatched BC does not imply that the BC is inappropriate, a matched BC is likely to be more accurately timed as the COPY strategy was rated as the most natural strategy. We consider a generated BC matching if there is a BC in the corresponding fragment of the COPY strategy whose onset is within a 200ms margin. This margin is strict but realistic given the precise timing that humans use in natural conversations [2]. The percentage of matched BCs can be regarded as the precision of the BC strategy, where the BCs in the COPY strategy are considered ground truth. Results are summarized in Table 1. Clearly, there are large differences between strategies. We note in particular the high precision of the PITCH&PAUSE strategy. Despite the lower number of generated BCs, the absolute number of matching BCs is higher than in the RANDOM strategy (10 and 6, respectively). From these observations, we conclude that both the number and timing of BCs contribute to the level of perceived naturalness.

Of note is also the relatively high precision of the RANDOM strategy. In the RANDOM and COPY strategies there was a BC every 6.77s and 6.24s, respectively. The average matching precision and recall between two uniform random distributions with the probability of observing a BC every 6.5s and a window size of 400ms (200ms in both directions) are 6.15%. The actual precision and recall of the RANDOM strategy are a factor two higher. We are left to conclude that the Erlang distribution that we used to model BCs in the RANDOM strategy is a good approximation of natural BC timings and/or that the timing of the specific set of generated BCs in the RANDOM strategy is exceptionally accurate.

Despite the higher number of BCs in the WARD&TSUKAHARA strategy, the score is much lower compared to the PITCH&PAUSE strategy. We observed that the matching precision of the WARD&TSUKAHARA strategy was low, which leads us to believe that the timing was less accurate. A possible explanation could be that Ward and Tsukahara [8] developed their strategy for audio-only settings and these might be systematically different from face-to-face settings when looking at speech and pitch features. This issue requires further research.

In an attempt to explain the low score for the GAZE strategy, we checked whether BCs would be systematically earlier or later compared to the COPY strategy. We found three times more matching BCs when the final rule of the GAZE strategy (Algorithm 2) was left out (corresponding to an overlap precision of 12.00%). This would also affect the PITCH,PAUSE&GAZE strategy. Again, there is no guarantee that the timing of the modified strategy will result in higher perceived naturalness. For the WARD&TSUKAHARA strategy, we did not find such a systematic bias.

Type of BCs. An important variable that we did not control for in our experiment was the BC type (nod, vocalization or combination of both). We take a closer look at the influence of type on the score. From Table 1, we see that there are some differences in the ratios between strategies, caused by the random factor in the generation of the BCs. We calculated the correlation between score and the percentage of vocalizations per fragment and strategy. We found a significant correlation ($r(94) = 0.201$, $p < 0.05$) which indicates that a higher percentage of vocals was found to be more natural. A fragment with only vocalizations on average scored almost 10 points higher than a fragment without vocalizations. This is somewhat at variance with remarks made by the participants. These reported that vocalizations placed within the speaker's discourse were disruptive and consequently rated lower. However, they also mentioned that vocalizations produced at appropriate moments gave them the impression that the BC behavior was performed by a human listener.

Participants also remarked that they found it unlikely that human listeners would nod when the speaker was not looking at them. It would therefore make sense to use information about the speaker's gaze also in the decision which BC type to animate.

Apart from the difference between individual BC types, the ratios of BC types performed by the actual listener are different from those reported in Dittmann and Llewellyn [9]. In a similar setting with one speaker and one listener they found approximately 60% of the observed BCs were vocalizations alone and around 22% were combinations of a nod and a vocalization. In contrast, we observed around 4% vocalizations alone and 30% combined BCs. One reason for the lower number of vocalizations in our experiment is that we did not select fragments with interjections that overlapped with the speaker's speech. Differences in familiarity, topic and conversation length might also have contributed to the discrepancy. Overall, the participants might have judged the BC behavior less natural due to potential differences in the ratio of BC type that we used and the ratio that is common for the type of conversation we considered in this experiment. However, this was not reported by any of the participants.

Role of blinks. Another factor in our stimuli was the presence of blinks. For each fragment and strategy, we animated blinks at the exact same moments as where the actual listener blinked. Especially for the fragments where no BCs were generated, this made the participants feel that they were still looking at an attentive listener. Several participants explicitly mentioned that they considered some blinks as BCs. Further research is needed to determine in which contexts blinks can be regarded as BCs, and whether they can be exchanged with different BC types.

6 Conclusions and Future Work

We have evaluated rule-based backchannel (BC) strategies in face-to-face conversations with the goal of improving BC behavior for artificial listeners. To this end, we used six different strategies to determine BC timings using features of the speaker's speech and gaze. We animated the BC behavior of an artificial listener for each of the six strategies, for a set of 16 fragments of recorded conversations. Given the multimodal nature of face-to-face conversations, we animated either a visual (nod) or vocal ("uhhuh") BC. Our stimuli consisted of a video of a human speaker and an animation of the artificial listener, shown side by side. In a user experiment, we had participants rate the likeliness that the BC behavior performed by the artificial listener had been performed by a human listener.

It appeared that a higher number of generated BCs increases the naturalness of the BC behavior. In addition, timing proved important. In fact, the strategy where the BC timings were identical to those in the original listener's video was rated the best. A third important yet uncontrolled factor was the type of animated BC. A positive correlation was found between the percentage of vocalizations and the rating. However, several participants in the experiment indicated that they thought it was unlikely that a human listener would produce vocalizations during the speaker's discourse. In our experiment, the use of the speaker's gaze did not result in more human-like BC behavior.

We believe that successful BC strategies in face-to-face conversations should accurately place BCs at a number of key moments. This is exactly what the PITCH&PAUSE strategy does. In addition, there should be a sufficient number of BCs throughout the speaker's turn. The PITCH&PAUSE strategy could therefore be combined with an Erlang model as used in the RANDOM strategy, or the algorithm of Ward and Tsukahara [8]. Finally, we propose to use keyword spotting (e.g. as in [24]) to respond immediately to acknowledgement questions such as "you know?" and "right?".

Future work will be focused on three aspects. First, we are currently working on a corpus-based evaluation, which allows to validate much more data and consequently overcome a potential bias due to the selection of a limited number of fragments [25]. This comes at the cost of a less precise evaluation as a high correlation with the actually performed BCs does not guarantee a high level of naturalness. We are looking at ways to combine the corpus-based evaluation with the annotation approach introduced in [11].

Second, we intend to look closer at the perception of animated BC types in different contexts. Insights in this direction could be used to determine not only when but also how to animate a BC. For example, to produce a visual BC only when the speaker looks at the listener. Finally, we aim to apply our strategies online, to generate BCs for an artificial listener in conversation with a human speaker.

References

1. Bavelas, J.B., Coates, L., Johnson, T.: Listeners as co-narrators. Journal of Personality and Social Psychology 79(6), 941–952 (2000)
2. Duncan Jr., S.: On the structure of speaker-auditor interaction during speaking turns. Language in Society 3(2), 161–180 (1974)
3. Yngve, V.H.: On getting a word in edgewise. In: Papers from the Sixth Regional Meeting of Chicago Linguistic Society, pp. 567–577. Chicago Linguistic Society (1970)
4. Bertrand, R., Ferré, G., Blache, P., Espesser, R., Rauzy, S.: Backchannels revisited from a multimodal perspective. In: Proceedings of Auditory-visual Speech Processing, Hilvarenbeek, The Netherlands, pp. 1–5 (August 2007)
5. Gravano, A., Hirschberg, J.: Backchannel-inviting cues in task-oriented dialogue. In: Proceedings of Interspeech, Brighton, UK, pp. 1019–1022 (September 2009)
6. Noguchi, H., Den, Y.: Prosody-based detection of the context of backchannel responses. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP), Sydney, Australia, pp. 487–490 (November 1998)
7. Okato, Y., Kato, K., Yamamoto, M., Itahashi, S.: Insertion of interjectory response based on prosodic information. In: Proceedings of the IEEE Workshop Interactive Voice Technology for Telecommunication Applications, Basking Ridge, NJ, pp. 85–88 (1996)
8. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics 32(8), 1177–1207 (2000)
9. Dittmann, A.T., Llewellyn, L.G.: Relationship between vocalizations and head nods as listener responses. Journal of Personality and Social Psychology 9(1), 79–84 (1968)

10. Morency, L.P., de Kok, I., Gratch, J.: A probabilistic multimodal approach for predicting listener backchannels. Autonomous Agents and Multi-Agent Systems 20(1), 80–84 (2010)
11. Huang, L., Morency, L.-P., Gratch, J.: Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Toronto, Canada (to appear, 2010)
12. Duncan Jr., S.: Some signals and rules for taking speaking turns in conversations. Journal of Personality and Social Psychology 23(2), 283–292 (1972)
13. Dittmann, A.T., Llewellyn, L.G.: The phonemic clause as a unit of speech decoding. Journal of Personality and Social Psychology 6(3), 341–349 (1967)
14. Kendon, A.: Some functions of gaze direction in social interaction. Acta Psychologica 26(1), 22–63 (1967)
15. Bavelas, J.B., Coates, L., Johnson, T.: Listener responses as a collaborative process: The role of gaze. Journal of Communication 52(3), 566–580 (2002)
16. Cathcart, N., Carletta, J., Klein, E.: A shallow model of backchannel continuers in spoken dialogue. In: Proceedings of the Conference of the European chapter of the Association for Computational Linguistics, Budapest, Hungary, vol. 1, pp. 51–58 (2003)
17. Maatman, M., Gratch, J., Marsella, S.: Natural behavior of a listening agent. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 25–36. Springer, Heidelberg (2005)
18. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R.J., Morency, L.P.: Virtual rapport. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 14–27. Springer, Heidelberg (2006)
19. Heylen, D., Bevacqua, E., Tellier, M., Pelachaud, C.: Searching for prototypical facial feedback signals. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 147–153. Springer, Heidelberg (2007)
20. Granström, B., House, D., Swerts, M.: Multimodal feedback cues in human-machine interactions. In: Proceedings of the International Conference on Speech Prosody, Aix-en-Provence, France, pp. 11–14 (2002)
21. Valstar, M.F., McKeown, G., Cowie, R., Pantic, M.: The Semaine corpus of emotionally coloured character interactions. In: Proceedings of the International Conference on Multimedia & Expo, Singapore, Singapore (to appear, 2010)
22. Boersma, P., Weenink, D.: Praat: doing phonetics by computer. Software (2009), `http://www.praat.org`
23. Van Welbergen, H., Reidsma, D., Ruttkay, Z., Zwiers, J.: Elckerlyc - A BML realizer for continuous, multimodal interaction with a virtual human. Journal of Multimodal User Interfaces (to appear, 2010)
24. Jonsdottir, G.R., Gratch, J., Fast, E., Thórisson, K.R.: Fluid semantic backchannel feedback in dialogue: Challenges and progress. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 154–160. Springer, Heidelberg (2007)
25. Truong, K.P., Poppe, R., Heylen, D.: A rule-based backchannel prediction model using pitch and pause information. In: Proceedings of Interspeech, Makuhari, Japan (to appear, 2010)