Auton Agent Multi-Agent Syst (2013) 27:235–253
DOI 10.1007/s10458-013-9219-z

Perceptual evaluation of backchannel strategies for artificial listeners

Ronald Poppe · Khiet P. Truong · Dirk Heylen

Published online: 13 January 2013

Abstract Artificial listeners are virtual agents that can listen attentively to a human speaker in a dialog. In this paper, we present two experiments where we investigate the perception of rule-based backchannel strategies for artificial listeners. In both, we collect subjective judgements of humans who observe a video of a speaker together with a corresponding animation of an artificial listener. In the first experiment, we evaluate six rule-based strategies that differ in the types of features (e.g. prosody, gaze) they consider. The ratings are given at the level of a speech turn and can be considered a measure for how human-like the generated listening behavior is perceived. In the second experiment, we systematically investigate the effect of the quantity, type and timing of backchannels within the discourse of the speaker. Additionally, we asked human observers to press a button whenever they thought a generated backchannel occurrence was inappropriate. Both experiments together give insights in the factors, both from an observation and generation point-of-view, that influence the perception of backchannel strategies for artificial listeners.

Keywords Backchannels · Listening behavior · Perceptual evalution · Artificial listeners

The research leading to these results has received funding from the European Community's 7th Framework Programme under Grant agreement 211486 (SEMAINE) and 231287 (SSPNet). Preliminary versions of this paper appeared as [29] and [30].

R. Poppe (✉)· K. P. Truong · D. Heylen
University of Twente, Human Media Interaction Group,
PO Box 217, 7500 Enschede, AE, The Netherlands
e-mail: r.w.poppe@utwente.nl

K. P. Truong
e-mail: k.p.truong@utwente.nl

D. Heylen
e-mail: d.k.j.heylen@utwente.nl

1 Introduction

Listening is an important aspect of conversation. In a dialog, the listener actively contributes to the conversation by signalling attention, interest and understanding to the speaker [1]. One particular type of signal is the *backchannel* [37], a short visual (e.g. nod, smile) or vocal (e.g. "uh-huh" or "yeah") signal from the listener that does not interrupt the speaker's speech and is not aimed at taking the turn. There are several types of backchannels [13]. Here, we focus on those with a *continuer* function that convey continued attention but carry no additional affective meaning. From the analysis of recorded human-human conversations, much is known about the type of backchannels and their placement within the discourse. We discuss these findings in Sect. 2.1.

Our goal is to use this knowledge to develop *artificial listeners*, virtual agents that can listen attentively to a human speaker [17]. This requires reliable prediction of backchannel opportunities from observations of the speaker's nonverbal visual and vocal behavior. In addition, appropriate listening behavior needs to be generated, which includes choosing the proper type of backchannel, and making sure that the number and spread of backchannels over a certain period of time is human-like.

Several authors have addressed real-time prediction of backchannel timings in audio-only settings using machine learning [26,27] or rule-based algorithms [35]. Due to the difficulty of analyzing lexical structure of the discourse in real-time, both typically rely on features that can be obtained with low processing requirements such as pitch slopes and pauses in the speaker's speech (e.g. [16]). Given certain values or changes in these features, the machine learning model or rule-based algorithm gives a (confidence) score that the production of a backchannel for the artificial listener is appropriate at the predicted moment.

Machine learning algorithms can automatically extract decision rules from features extracted 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 [35] manually defined a rule-based algorithm that predicts backchannels based on the speaker's pitch contours. They obtained reasonable results while the algorithms are easy to understand and verify. In this paper, we focus on rule-based algorithms for these reasons.

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 [11] observed that the mere fact that conversational partners can see each other is a signal of attention and thus reduces the need for backchannels. Also, the turn-taking process, including backchannel feedback, is arguably more complex. Duncan [13] identifies speech, gaze and gesture as relevant components of natural turn-taking, while turn-taking in audio-only settings only considers speech.

Typically, algorithms that predict backchannel timings have been evaluated using corpus-based measures such as precision and recall, which are informative of how well the prediction matches the backchannels that have actually been performed in the corpus. However, good approximation of backchannel timings is not a guarantee that the predicted behavior will be *perceived* as human-like. This is partly due to the optionality of backchannels. For example, a predicted backchannel that is not performed by the human listener in the corpus is not necessarily incorrect, and vice versa. Moreover, other factors such as the type of backchannel and the number of backchannels in a period of time are not taken into account in corpus-based research. As our aim is to develop artificial listeners that are to be perceived as human-like, we therefore evaluate the generated listener behavior using perceptual measures.

In this paper, we present two experiments where we investigate the perception of rule-based backchannel strategies for artificial listeners. Both rely on subjective judgements of humans who observe at the same time a video of a speaker and a corresponding animation of an artificial listener. In the first experiment, we evaluate six rule-based strategies that differ in the types of features (e.g. prosody, gaze) they consider. The ratings are given at the level of a speech turn and can be considered a measure for how human-like the generated listening behavior is perceived. In the second experiment, we systematically investigate the effect of the quantity, type (visual, vocal) and timing of backchannels within the discourse of the speaker. In addition to turn-level subjective ratings, we asked human observers to press a button whenever they thought a generated backchannel occurrence was inappropriate. Both experiments together give more insights in the factors, both from an observation and generation point-of-view, that influence the perception of backchannel strategies for artificial listeners.

The remainder of this paper is organized as follows. We discuss related work on listening behavior, artificial listeners and the evaluation of backchannel strategies in the next section. The six backchannel strategies that we will evaluate are described in Sect. 3. Section 4 summarizes the data used for our stimuli. Experiments 1 and 2 are presented in Sects. 5 and 6, respectively. Finally, we discuss our results and present promising directions for future research.

2 Related work

We first discuss related literature on listening behavior, with a specific focus on backchannels. We then turn to artificial listeners, and summarize the various efforts that have been undertaken to endow virtual agents with listening capabilities. Finally, we describe the current practice of the evaluation of backchannel strategies, specifically for artificial listeners.

2.1 Listening behavior

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 backchannel feedback was first used by Yngve [37], who described it as messages sent by the listener without the intent to take the turn. Research into backchannels can be grouped into two directions [36]: the lumping approach and the splitting approach. The former treats backchannels as a single class and is mainly concerned with the timing within a speaker's discourse. The latter approach has investigated specific forms of backchannels and their role in a turn-taking context.

Backchannels take many forms including short vocalizations (e.g. "hmm", "uhhuh"), sentence completions, requests for clarification, brief restatements, and facial and bodily manifestations such as smiles [6] and head nods [13]. 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 backchannels in conversation, researchers have focused on identifying the nonverbal context of backchannels. Duncan [12] observed that backchannels are often used as a response to a speaker's turn-yielding signal, implying that they might be cued by the speaker. Dittmann and Llewellyn [10], Duncan [13] and Yngve [37] noted that backchannels are often produced after rhythmic units in the speaker's speech, and specifically at the end of grammatical clauses. A region of low or rising pitch [16], a high or decreasing energy pattern [23] and a short pause [7] in the speaker's speech have been found to precede backchannels from the listener. These prosodic markers often occur at end of grammatical clauses or at the end of the speaker's turn.

In face-to-face interactions, Kendon [21] and Bavelas et al. [2] looked at the relation between gaze and backchannels, and found that the listener is likely to produce a backchannel when there is a short period of mutual gaze between speaker and listener. These moments also usually occur at the end of a speaker's turn.

2.2 Artificial listeners

Our goal is to develop artificial listeners, virtual agents that can listen attentively to a human speaker. This requires analysis of the speaker's verbal and nonverbal behavior to identify moments where backchannels might be produced. The systematics in the placement of backchannels within the conversation have motivated researchers to train machine learning models or devise rule-based algorithms for the identification of backchannel opportunities. Recent work has been focused on using low-level features for the prediction or generation of backchannel timings. Machine learning models such as decision trees [26], Conditional Random Fields [8,25], Hidden Markov Models [27] and Support Vector Machines [9] have been used for backchannel timing generation. These models are usually unintuitive and require suitable training data. In contrast, rule-based models (e.g. [31,35]) are typically specified manually and are easy to understand.

In addition to the identification of backchannel opportunities, artificial listeners need to display human-like backchannel behavior. Attempts in this direction have been taken by Huang et al. [18], who generated head nods at moments that had been identified off-line based on multi-modal features. Maatman et al. [24] used the rule-based prediction algorithm of Ward and Tsukahara [35] and displayed nods in an online setting. Even though both works considered a face-to-face setting, none of them have addressed the type of backchannel.

2.3 Evaluation of backchannel strategies

Systems that automatically predict backchannel timings are usually evaluated using the correlation between the predicted and actually performed backchannels. This approach does not take into account individual differences. A given moment where an individual does not provide a backchannel is not necessary an inappropriate moment for backchannel feedback. However, in the evaluation of backchannel prediction algorithms, such a predicted backchannel would be regarded as a false positive.

Individual differences also affect the training of machine learning models as samples without backchannels are labeled as negatives which consequently decreases the quality of the model. This issue was addressed by Huang et al. [18], who had observers watch a video of a speaker and indicate where they would provide backchannels as if they were actual listeners. They analyzed which backchannel 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.

In the lumping approach, backchannels have been treated as a single class, without distinguishing between the type. However, there are systematic differences in the structural properties of backchannels of different types. For example, Dittmann and Llewellyn [11] observed that, on average, a nod is produced 175ms earlier than a vocalization. Truong et al. [32] found that a visual backchannel was more likely to occur during mutual gaze, whereas vocal backchannels were more often produced during a pause in the speaker's speech.

Similar findings have been reported in [3]. Given these differences in occurrence, it is likely that there is no such thing as a backchannel opportunity, but rather the opportunity for a specific type of backchannel. We therefore expect that a different type of backchannel, produced in the same structural context, will be perceived differently by human observers.

An evaluation of the perceived relative timing (e.g. too early) was performed by Kitaoka et al. [22] in audio-only settings. The perception of different types of backchannels has been researched in [4,14] but in isolation, not in a conversational context. The use of backchannels was one of a number of ways to increase the feeling of rapport with a virtual agent in [15,19]. However, no evaluation of different backchannel timing algorithms or characteristics of the algorithm were carried out.

In summary, a purely corpus-based evaluation does not seem appropriate for the evaluation of backchannel strategies. To this end, in this paper, we present two experiments that investigate how different rule-based backchannel strategies are perceived by human observers. We investigate the quality of the different strategies, and analyze the different factors in the generation of the backchannel behavior.

3 Backchannel strategies

We define six strategies to determine the placement of listener backchannels in real-time based on the speaker's speech, gaze or both. An artificial listener produces backchannels based on the identified timings. In the experiments presented in this paper, we use fragments from a corpus (see Sect. 4) with human speakers and listeners, and we present participants with video samples in which the human listener is replaced by an artificial listener. However, all strategies are aimed at use in an online dialog. We discuss the strategies below.

- **Copy** This strategy produces a backchannel at those moments where the actual listener performed a backchannel in interaction with the speaker. We include this strategy to investigate whether the acutally performed backchannel behavior is perceived as more appropriate.
- **Random** An approximation of an Erlang distribution is used to generate backchannel timings without taking into account any signal from the speaker. We use one normal distribution to model the timing of the first backchannel, and one to model the time between two subsequent backchannels. For the generation of backchannel onsets, we iteratively sample from the distributions until the end of the fragment is reached. We resample when the time between subsequent backchannels is below 1 s. One random distribution for each strategy-fragment combination is generated.
- **Ward & Tsukahara** We use the rule by Ward and Tsukahara [35] that has been used for backchannel prediction in English audio-only settings, reprinted as Algorithm 1.

Algorithm 1: WARD & TSUKAHARA strategy for backchannel placement

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

- **Gaze** Several researchers have observed the relation between (mutual) gaze and backchannels. In a setting similar to ours, Bavelas et al. [2] 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 backchannel is likely to occur. Similar observations have been made by Duncan [12] and Kendon [21]. We use this mechanism to determine the timing of backchannels based on the speaker's gaze at the listener only, formalized in Algorithm 2.

Algorithm 2: GAZE strategy for backchannel placement

```
Provide backchannel feedback upon detection of:
P1 gaze at the listener,
P2 coming after at least 1,000 ms of no gaze,
P3 after 500 ms wait.
```

- **Pitch & Pause** It has been observed that backchannels frequently occur in a pause after a speaker's utterance [10,37]. This observation has led to the introduction of Algorithm 3 [31]. Similar to Truong et al. [31], we use a minimum pause duration of 400 ms as a compromise between the 200 ms as used in Maatman et al. [24] and the 700 ms used by Ward and Tsukahara [35]. We further take into account the preceding speech. Instead of a region of low pitch, we focus on rising or falling pitch, as suggested by several researchers [11]. Gravano and Hirschberg [16] found in their audio-only corpus that over 80 % of all backchannels were preceded by either a rising or falling pitch contour.

Algorithm 3: PITCH & PAUSE strategy for backchannel placement

```
Provide backchannel feedback upon detection of:
P1 a pause of 400 ms,
P2 preceded by at least 1,000 ms of speech,
P3 where the last 100 ms,
P4 contain a rising or falling pitch of at least 30 Hz.
P5 provided that no backchannel has been output within the preceding 1,400 ms.
```

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

4 Stimuli

We used the Semaine Solid SAL data [33], which contains dialogs between a human listener and a human speaker. The human listeners were instructed to play four different characters: an optimistic, a depressed, an angry and a pragmatic person. Given that we are not so much interested in the emotional overtones of the listener's reactions we only consider interactions with Prudence, the pragmatic character. From these, we selected fragments where the character was played by two different listeners.

We automatically extracted speaker fragments bounded by 1.5 s of non-speech of the speaker. We discarded fragments that were shorter than 10 s, did not contain backchannels from the listener or that contained interjections (e.g. "that's good") that overlapped with the speech of the speaker. From the remaining fragments, we selected eight samples for each of the two listeners. We further removed the listener's vocalizations from the speaker's audio signal, if needed. The average sample length was 19.9 s, with an average of 2.6 backchannels per fragment.

Speaking/pause and pitch information were obtained using Praat [5], speaker's gaze towards the listener was annotated manually. Alternatively, we could have used a gaze tracker. We annotated the onset of all backchannels (nods and vocalizations) from the listener's video. In the case of repeated nods, we took the most articulated nod. When two backchannels overlapped in time (e.g. nod and vocal), we annotated a single instance with the onset of the earliest backchannel. In this study, we did not consider smiles and other facial signals as backchannels as these typically serve a more specific function in the discourse.

In our stimuli, we replaced the video of the listener by a virtual agent. We used Elckerlyc [34], 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 chose two common generic (see [1]) backchannels in face-to-face interaction: a nod and vocalization ("uhhuh"). There is surprisingly little known about potential semantic differences between visual and vocal backchannels. Duncan [13] found no difference in the placement within the speaker's discourse. This was also observed by Dittmann and Llewellyn [11], who did note that a nod on average precedes a vocalization by 175 ms. However, such a temporal difference was not observed in an analysis in [32]. The large number of backchannels in our sample sets did not allow for a controlled design and we introduced backchannel type as an uncontrolled variable in Experiment 1. At each defined backchannel onset, we animated a nod, a vocalization or a combination of both with the same ratio as performed by the actual listeners, calculated over all fragments. The nods were relatively subtle and had a duration of 400 ms. In Experiment 2 (Sect. 6), we will investigate how the type of backchannel influences the perception of the listening behavior.

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 backchannels, but it is unclear to what extend they can be replaced by a different type of backchannel. 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 Fig. 1). It should be noted that all other behaviors of the artificial listener, including head and body movement, gaze behavior and facial expressions, were not varied. While this renders the artificial listener somewhat unnatural, it ensures that effects found are not due to differences in uncontrolled behaviors.

5 Experiment 1: perception of backchannel strategies

The aim of this experiment is to investigate the quality of the six backchannel strategies introduced in Sect. 3. On the one hand, we use objective evaluation metrics that compare the generated backchannel timings of each strategy to those performed by the actual listener in the corpus. On the other hand, we use subjective measures obtained from human observers

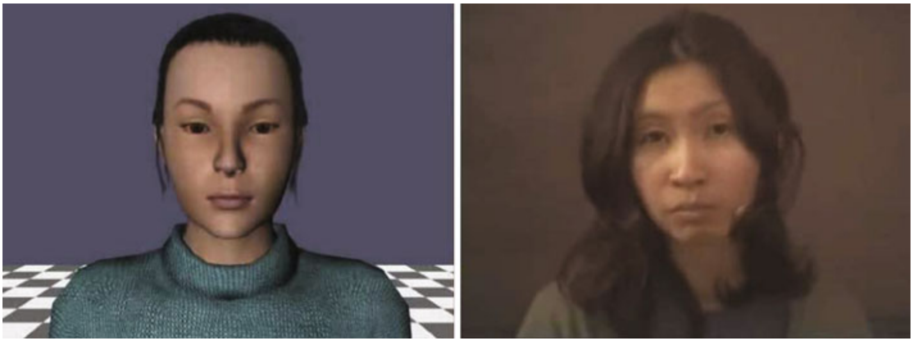


Fig. 1 Example stimulus with artificial listener (*left*) and actual speaker (*right*)

that viewed the stimuli. We asked them to rate how likely they thought the displayed listening behavior was performed by a human listener. We discuss the details of the experiment below, and present the results in Sect. 5.4.

5.1 Stimuli

We used the stimuli as described in Sect. 4. The six backchannel strategies were applied to the 16 fragments, resulting in a total of 96 stimuli.

For the RANDOM strategy, the mean (SD) start of the first backchannel was at 6.97 s (6.20 s), with 5.00 s (2.73s) between subsequent backchannels.

5.2 Procedure

It was explained to the participants that they would be participating in an experiment to determine the quality of backchannel 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, after viewing a 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 listeners, performed with all six backchannel 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 variables and order as between-subjects variable.

5.3 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.

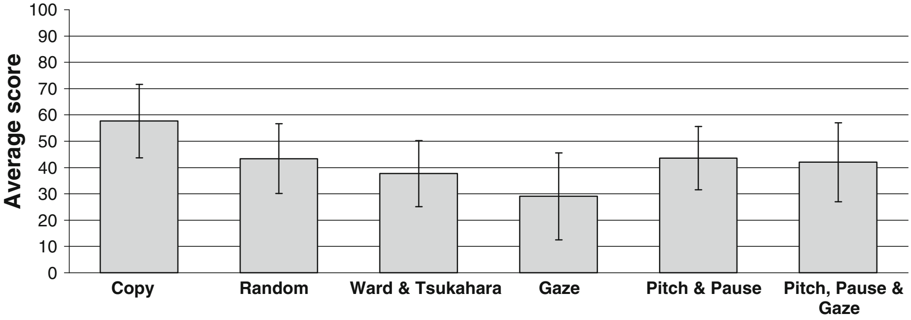


Fig. 2 Average scores (with SD) for all strategies, calculated over all responses

Table 1 Summary of results per 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 backchannels	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

A generated backchannel is matching if its onset appears within a margin of 200 ms of the onset of a backchannel in the COPY strategy

5.4 Results and discussion

The 20 participants rated in total 960 samples. A session length was approximately 20–25 min. We found no significant effect for the number of the stimulus, and we expect that neither learning nor fatigue have had an important effect. 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 strategy variable proved significant (F(5, 80) = 22.141, $p < 0.01$). See Fig. 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 standard deviations for each strategy are partly explained by differences in rating scores of individual participants. While the average score over all samples was approximately 42, these average scores ranged between 17 and 62 for individual participants. An analysis with (normalized) z-scores resulted in the same interaction effects. In the following, we will use the scores without normalization.

Overall, the scores are rather low. This might be partly because many behaviors that typically play a role in listening behavior were not animated. This was mentioned several times by the participants. 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 backchannels was performed by the actual listener, although in many cases the animated backchannel type was different from the type that was actually performed. The relatively low score of the COPY condition might partly be attributed to the backchannel type. We used two generic backchannel types, which might not be the most suitable choice in all cases. In an engaged conversation, a smile might have been a better choice in some cases. Also, we expect that part of the lower score can be attributed to inter-personal differences in backchannel behavior. Several participants reported that they performed backchannels based on the speaker's video as if they were the listener. Simultaneously, they monitored the animation of the listener to see how well the animated backchannels corresponded with their own backchannels. Inter-personal differences in the production of backchannels can be significant [8].

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. We now investigate several factors that might explain these findings.

5.4.1 Number of backchannels

In a first attempt to explain these differences, we analyzed the effect of the number of backchannels. We calculated the correlation between the number of backchannels 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 backchannels per minute was between 0 and 23. Given this correlation, it is more likely that strategies with a lower number of generated backchannels 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 backchannels was generated (see Table 1). Also, the additional backchannels of the GAZE strategy did not increase the score of the PITCH, PAUSE & GAZE strategy compared to the PITCH & PAUSE strategy. A systematic analysis of the number of backchannels will be carried out in Experiment 2 (Sect. 6).

5.4.2 Timing of backchannels

To quantitatively investigate the timing of the backchannels, we calculated how well these matched the backchannels in the COPY strategy. While an unmatched backchannel does not imply that the backchannel is inappropriate, a matched backchannel is likely to be more accurately timed as the COPY strategy was rated as the most natural strategy. We consider a generated backchannel matching if there is a backchannel in the corresponding fragment of the COPY strategy whose onset is within a 200 ms margin. This margin is strict but realistic given the precise timing that humans use in natural conversations [13,32]. The percentage of matched backchannels can be regarded as the precision of the backchannel strategy, where the backchannels 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 backchannels, the absolute number of matching backchannels is higher than in the RANDOM strategy (10 and 6, respectively). From these observations, we conclude that both the number and timing of backchannels contribute to the level of perceived naturalness. It should be noted, however, that these percentages have been calculated over small numbers of backchannels. We will further investigate this in Experiment 2 (Sect. 6).

Of note is also the relatively high precision of the RANDOM strategy. In the RANDOM and COPY strategies there was a backchannel every 6.77 and 6.24 s, respectively. The average matching precision and recall between two uniform random distributions with the probability of observing a backchannel every 6.5 s and a window size of 400 ms (200 ms in both directions) are both 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 backchannels in the RANDOM strategy is a good approximation of natural backchannel timings and/or that the timing of the specific set of generated backchannels in the RANDOM strategy is exceptionally accurate.

Despite the higher number of backchannels 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 AND TSUKAHARA strategy was low, which leads us to believe that the timing was less accurate. A possible explanation could be that Ward and Tsukahara [35] 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 backchannels would be systematically earlier or later compared to the COPY strategy. We found three times more matching backchannels 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 AND TSUKAHARA strategy, we did not find such a systematic bias.

5.4.3 Type of backchannels

An important variable that we did not control for in our experiment was the backchannel 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 backchannels. 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 backchannel behavior was performed by a human listener. In Experiment 2 (Sect. 6), we will take a close look at the influence of type on the perception.

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 backchannel type to animate.

Apart from the difference between individual backchannel types, the ratios of backchannel types performed by the actual listener are different from those reported in Dittmann and Llewellyn [11]. In a similar setting with one speaker and one listener they found approximately 60 % of the observed backchannels 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 backchannels. 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 backchannel behavior less natural due to potential differences in the ratio of backchannel 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.

5.4.4 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 backchannels 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 backchannels. Further research is needed to determine in which contexts blinks can be regarded as backchannels, and whether they can be exchanged with different backchannel types.

6 Experiment 2: effect of quantity, type and timing

In Experiment 1, we did not control for the quantity, type and timing of the generated backchannels. We found indications that these factors influence the perception of the backchannel behavior. In this experiment, we systematically vary these factors and measure the effect on the perception.

From Experiment 1, we expect a similar significant positive correlation between the number of backchannels and the rating of the fragment. We again use two types of backchannels: visual (nod) and vocal ("uh-huh"). While both types can have the same continuer function, there are differences in timing within the speaker's turn. For example, nods are more often produced during mutual gaze, whereas vocalizations tend to be produced around the end of a segment of speech [32]. With regards to specific timing, we therefore expect that there is no such thing as a general backchannel opportunity, but rather an opportunity for a nod or an opportunity for a vocalization. Although both probably partly overlap, in general, we expect that changing the type from that was actually performed would result in lower subjective ratings. While backchannels are optional, there are many known systematics in the production of a backchannel as a reaction or anticipation of the speaker's verbal and nonverbal behavior. We expect that contingent timings, rather than random timings, will be rated as more human-like.

An important addition to the experiment procedure was that participants were not only asked to give a rating per fragment, but also to judge individual backchannels. A common observation in the area of virtual agents is that humans are sensitive to the flaws in animated behavior. With this in mind, we introduced the *yuck* button approach: while watching the stimuli, a button is pressed every time a human observer thinks the behavior displayed is inappropriate. This approach allows us to obtain subjective ratings for both fragments and individual backchannels without additional time requirements. In turn, we can analyze how the rating of individual backchannels influences the perception of an entire fragment.

6.1 Stimuli

We selected the 12 fragments from Sect. 4 that had at least two backchannels. The fragments are between 14 and 31 seconds in length. The backchannel behavior of the artificial was systematically varied along three dimensions: quantity, type and timing. For each dimension, we took the manually annotated backchannels performed by the actual listener as a basis. This corresponds to the COPY condition of Experiment 1. For the quantity dimension, we defined three conditions. All backchannels were used in the *original* condition. In the *odd* and *even* condition, we selected every second backchannel, starting with the first or second one, respectively. The three conditions contained 46, 26 and 20 backchannels, respectively. We used the same backchannel types as in Experiment 1: a nod, a vocalization ("uh-huh") or a combination of both. We animated either the *original* types, or the *switched* types, with nods replaced by vocalizations, and vocal and bimodal backchannels by nods. To test whether backchannels should be placed at specific moments within the discourse, we used the *original* onsets or *random* onsets for the timing dimension. The latter ones where determined by sampling an Erlang distribution as in the RANDOM condition of Experiment 1. The order of the types of backchannels was left unchanged. The three dimensions were crossed to yield 12 conditions. In addition to the backchannels, we animated the listener's blinks where they occurred in the actual recording.

6.2 Procedure

Again, it was explained to the participants that they would be participating in an experiment to determine the quality of backchannel behavior. After the briefing, they were shown a set of stimuli. They were instructed to press the yuck button (space bar) every time they thought a listener's backchannel was inappropriate, either in type or timing. Participants could replay the video as often as desired, and adapt their yucks if needed. After watching a video fragment, they were prompted to rate how human-like they perceived the listener's backchannel behavior as in Experiment 1.

We divided the 144 condition-fragment combinations into six distinct sets of 24 stimuli. We adopted a Greco-Latin square design to control for order. In addition, this ensured that each participant rated each fragment and condition twice and six participants together rated all possible combinations of both.

6.3 Participants

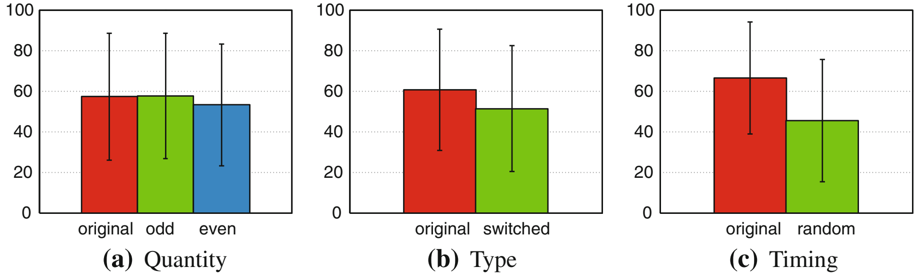
We recruited 24 colleagues and doctoral students (6 female, 18 male) with a mean age of 32.8 (min 23, max 58). Each of the participants was assigned randomly to a set of stimuli with the lowest number of respondents.

6.4 Results and discussion

We collected ratings over fragments, and yucks for individual backchannels. Both are discussed separately in the following sections.

6.4.1 Fragment ratings

We analyze how quantity, type and timing of backchannels affects how human-like participants perceived a fragment. Similar to Experiment 1, we ignore the variable fragment as the number of ratings per fragment-condition combination is limited. We performed a repeated measures ANOVA with order as between-subjects variable and quantity, type and timing as within-subjects variables. There are differences in ratings for different participants. While these do not affect the significance of the observed effects, we will use z-scores of the fragment ratings as our dependent variable unless explicitly stated otherwise.



(a) Quantity **(b)** Type **(c)** Timing

Fig. 3 Average un-normalized fragment ratings per dimension

In Fig. 3, the results of un-normalized ratings for quantity, type and timing are shown. Overall, and similar to the fragment ratings obtained in Experiment 1, these scores are rather low. We again attribute this observation to the fact that only backchannels and blinks were animated. The high standard deviations are due to the grouping of ratings from different fragments.

For quantity, we did not find a significant effect (F(2)=2.062, $p = ns$). We observe in Fig. 3a that the difference between 46 and 26 backchannels in the *original* and *odd* conditions is minimal. However, there is an interaction effect between quantity and timing (F(2)=5.891, $p < 0.01$). Specifically, the effect is opposite; the *odd* condition was rated the best with the *original* timings, but lowest with *random* timings. The differences between the quantity conditions for the *random* timings are less pronounced.

In this analysis, we did not account for the duration of the fragment. If we correlate the average ratings per fragment-condition combination with the average number of backchannels per minute, we find no significant effect. More backchannels per minute are thus not perceived as more human-like. However, when only the conditions with *original* timing and type are taken into account, the correlation is significant (r(36)=0.340, $p < 0.05$). Closer analysis reveals that the fragment with the highest number of backchannels per minute (20.18) appears to be an outlier. Leaving this fragment out results in a correlation of r(35)=0.537, $p < 0.001$. This analysis suggests that too few and too many backchannels will reduce the quality of the backchannel behavior. We expect that a reasonable number of backchannels per minute lies somewhere between 6 and 12.

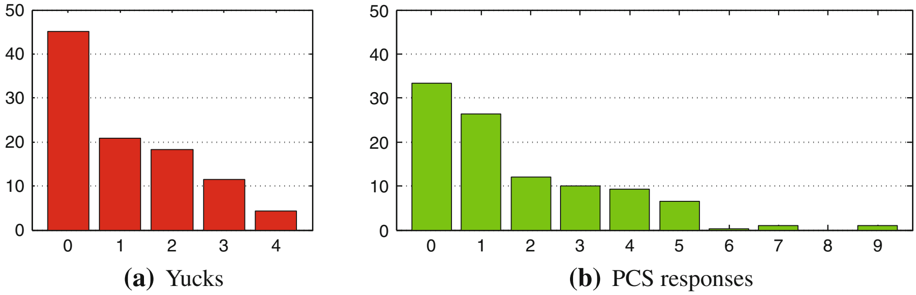
Type proved to be significantly different for the *original* and *switched* conditions (F(1)=18.233, $p < 0.001$). Apparently, different types of backchannels are performed in different contexts. We will investigate this more thoroughly in Sect. 6.4.2.

We also found a main effect for timing (F(1)=94.684, $p < 0.001$). Apparently, participants rated *random* timing lower. Original timings are perceived as more human-like. However, the difference between the two conditions is moderate. A similar observation was also made in Experiment 1 and in [18] and can be partly attributed to inter-personal differences, the optional nature of backchannels and the fact that, apart from backchannels and blinks, no other behaviors were animated.

While these results reveal differences in perception for different quantity, type and timing conditions, each fragment contains multiple backchannels. In the next section, we will analyze the perception of individual backchannels.

6.4.2 Individual backchannel ratings

For each backchannel, we are interested in how often participants rated it as inappropriate. We obtain this information by linking the yucks to the performed backchannels. In addition, we



(a) Yucks

(b) PCS responses

Fig. 4 Relative frequency of yucks and PCS responses per backchannel (%)

obtain a measure of optionality for each backchannel using parasocial consensus sampling (PCS), which we explain next.

Given that backchannels are often optional and that there are inter-personal differences in backchannel behavior, we are interested in the optionality of specific backchannels. We used the concept of PCS [18] as a means to obtain backchannel opportunities from multiple raters. Specifically, we had nine participants watch the video of the speaker from the fragments that were used in the perception experiment. We asked them to press a button whenever they would perform a backchannel. In total, we obtained 240 responses. We expect that the ratings give a more general idea at which moments backchannels are common, and when they are more optional. Next, we discuss how we linked the PCS and yuck responses to the backchannels generated in the stimuli.

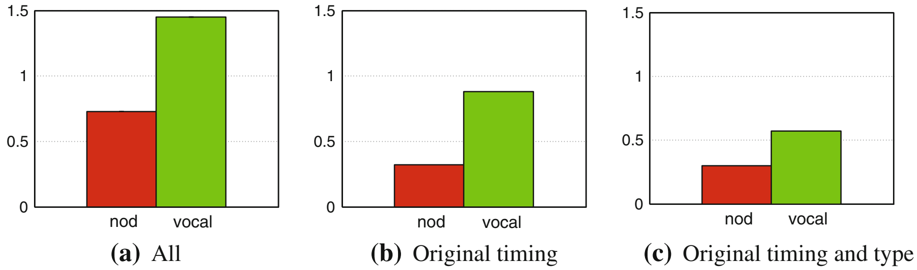
As our aim is to report on the appropriateness of individual backchannels in the stimuli, we need to associate the yucks and the PCS responses to these backchannels. For the yucks, there is a time delay between the stimulus onset and the participants' key press. We analyzed this delay and associated a yuck response with the closest preceding backchannel, provided that the time between them was between 300 and 2,500 ms.

One would expect that the timings of PCS responses are similar to the actual backchannel onsets. On closer analysis, a PCS response appears to be approximately 200 ms later. We use a matching window of 500 ms and therefore, we associate a PCS with a backchannel if it is between 300 ms before and 800 ms after a backchannel onset.

The total number of generated backchannels in all fragments and conditions is 368. Figure 4 shows the frequency of yucks and PCS responses per backchannel. Each fragment-condition combination has been judged by four participants, so the maximum number of yucks per backchannel is four. As the numbers of yuck and PCS responses are a measure of a backchannel's unsuitability and suitability, respectively, it is not surprising that the numbers of these responses are negatively correlated (r(368) = -0.400, $p < 0.001$).

Quantity For now, we only consider the quantity dimension, and use only the data of the *original* type and timing conditions. As the *odd* and *even* quantity conditions contain a subset of the backchannels in the *original* quantity condition, we expect similar numbers of PCS responses in all conditions. These numbers are 2.35, 2.58 and 2.06, respectively. They are reasonably equal and correlate with the fragment ratings.

If quantity would not be an important factor in backchannel behavior, we would expect similar numbers of yucks as well. However, we found the average numbers of yucks per backchannel to be 0.54, 0.19 and 0.25 for the *original*, *odd* and *even* conditions, respectively. Apparently, more backchannels is not always better. This is somewhat at variance with



(a) All **(b)** Original timing **(c)** Original timing and type

Fig. 5 Average number of yucks per nod or vocalization in different conditions

findings in Experiment 1. Closer analysis reveals that eight out of 25 yucks in the *original* setting originate from the fragment with the highest number of backchannels per minute (20.18). This fragment was also rated much lower in general (see Sect. 6.4.1). Again, it appears that too few or too many backchannels reduces the perceived quality of the backchannel behavior.

Type In Sect. 6.4.1, we investigated differences between the *original* and *switched* condition as an indication that changing the type of backchannel affects how it is perceived. Given the yucks, we can also analyze whether the type of an individual backchannel matters, regardless of the specific condition. We expect that the vocal aspect of bimodal backchannels is most salient as it might overlap with the speaker's speech. Therefore, we treated these backchannels as vocalizations. The average numbers of yucks for nods and vocalizations are, respectively, 0.32 and 0.88 with *original* timing, and 1.15 and 2.01 with *random* timing (see Fig. 5). Over both conditions, the percentage of backchannels that did not receive a yuck was 57.6 and 32.6 % for the nods and vocalizations, respectively. We can further narrow down the class with *original* timings and distinguish between the backchannels performed in the *original* and *switched* type conditions. Changing vocalizations and bimodal backchannels to nods caused a slight increase in number of yucks per backchannel, from 0.30 to 0.36. However, changing nods to vocalizations led to an increase from 0.57 to 1.02.

These numbers indicate that a nod is less often perceived as inappropriate. We expect this can at least be partly explained by the fact that nods are communicated over the visual channel, without directly interfering with the main channel of communication. Therefore, it might be that vocalizations are more precisely timed, whereas nods can be performed throughout the speaker's turn. If this would be the case, one would expect higher numbers of PCS responses for a vocalization compared to a nod for the actually performed backchannels. This is indeed the case, with on average 3.20 responses for a vocalization and 1.97 for a nod. These findings are important for the design of backchannel generation algorithms for artificial listeners. High confidence in the backchannel prediction could result in the production of a vocalization, whereas a nod might be produced otherwise.

Timing The effect of timing on the perception of backchannel behavior can also be observed from the PCS and yuck responses. The average number of PCS responses for a backchannel with *random* timing is 1.03, compared to 2.35 with the *original* timing. Randomly timed backchannels are thus twice less likely to occur. Not surprisingly, the number of yucks in the random condition is much higher than in the original condition, 1.58 versus 0.60. This again shows that timing matters.

7 Discussion and conclusions

We have evaluated rule-based backchannel strategies in face-to-face conversations with the goal of improving backchannel behavior for artificial listeners. To this end, we used six different strategies to determine backchannel timings using features of the speaker's speech and gaze. We animated the backchannel 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") backchannel. 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 backchannel behavior performed by the artificial listener had been performed by a human listener.

There appeared to be differences in the perceived quality of the different backchannel strategies. The backchannels generated at the moments where the actual listener performed them, were judged best. The PITCH & PAUSE strategy that generates backchannels in pauses after a rising or falling speaker pitch contour received moderate scores. Its extension to also include those moments where a period of mutual gaze starts scores on par. Gaze alone and the algorithm of Ward and Tsukahara [35] yielded the lowest fragment ratings. A random strategy that used an Erlang distribution but did not rely on features from the speaker's speech or gaze resulted in a moderate score as well.

Differences in the quantity, type and timing of generated backchannels appeared to affect these ratings as well. As these factors were not controlled, we performed a second experiment where we used the originally performed backchannels (COPY condition) and varied the quantity, type and timing of these in a controlled manner. In addition to the fragment ratings, we had human observers identify the backchannels they thought were inappropriate in the context of the discourse. These ratings allowed us to give quantitative results at the level of individual backchannels.

From the first experiment, it appeared that a higher number of generated backchannels increases the naturalness of the backchannel behavior. In Experiment 2, we observed that there is indeed a trend that a higher backchannel rate increases the perceptual ratings. However, there appears to be a ceiling effect. We expect that a reasonable number of backchannels per minute lies somewhere between 6 and 12. The type of generated backchannel was also of influence. In Experiment 2, switching the type of a backchannel resulted in more yucks. Apparently, different types of backchannels are performed in different contexts. Analysis of individual backchannels revealed that nods are less often rated as inappropriate, regardless of their timing. This knowledge has implications for the design of backchannel generation algorithms. If the prediction confidence is low, it is probably more appropriate to generate a nod. Thirdly, timing proved important. In both Experiment 1 and 2, the strategy where the backchannel timings were identical to those in the original listener's video (COPY condition) was rated the best.

We believe that successful backchannel timing generation algorithms in face-to-face conversations should accurately place backchannels at a number of key moments. This is exactly what the PITCH & PAUSE strategy does. In addition, there should be a sufficient number of backchannelss throughout the speaker's turn. An average between 6 and 12 backchannels per minute appears to be suitable. 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 [35]. This should ensure that the time between two backchannels is realistic. Finally, we propose to use keyword spotting (e.g. as in [20]) to respond immediately to acknowledgement questions such as "you know?" and "right?".

In addition, when generating a backchannel, we should take into account its type. This is true in general, as different types of backchannels can carry different meanings, but it also holds in the case which we investigated in which we used backchannels from different modalities with a continuer function. When the confidence of the backchannel timing algorithm is low, a nod is more suitable as it is less often perceived as inappropriate. When the confidence is high, an accurately timed vocalization is likely to be perceived as more human-like backchannel behavior.

The presented experiments have been performed offline, on recorded video fragments. Future work should address online settings, with an artificial listener in conversation with a human speaker. Such a setting requires that we are able to evaluate the backchannel strategies in real-time based on features observed from the speaker's speech and gaze, and potentially other modalities. In [28], a method is presented to evaluate the generated backchannel behavior online using a yuck button approach, in a similar manner as we did in Experiment 2. Future work will be aimed at online evaluation of backchannel timing algorithms for artificial listeners that can be continuously adapted (e.g. as in [9]), to ensure that the performed behavior will be perceived as being human-like.

Acknowledgments The authors would like to thank Dennis Reidsma for his assistance in preparing the stimuli, and Iwan de Kok for interesting discussions.

References

1. Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology*, *79*(6), 941–952.
2. Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, *52*(3), 566–580.
3. Bertrand, R., Ferré, G., Blache, P., Espesser, R., & Rauzy, S. (2007). Backchannels revisited from a multimodal perspective. *Proceedings of Auditory-Visual Speech Processing* (pp. 1–5). Hilvarenbeek, The Netherlands.
4. Bevacqua, E., Pammi, S., Hyniewska, S.J., Schröder, M., & Pelachaud, C. (2010). Multimodal backchannels for embodied conversational agents. *Proceedings of the International Conference on Interactive Virtual Agents (IVA)* (pp. 194–200). Philadelphia.
5. Boersma, P., & Weenink, D. (2009). Praat: Doing phonetics by computer. Software. www.praat.org.
6. Brunner, L. J. (1979). Smiles can be back channels. *Journal of Personality and Social Psychology*, *37*(5), 728–734.
7. Cathcart, N., Carletta, J., & Klein, E. (2003). A shallow model of backchannel continuers in spoken dialogue. *Proceedings of the Conference of the European chapter of the Association for Computational Linguistics*, vol. 1 (pp. 51–58). Budapest, Hungary.
8. de Kok, I., Ozkan, D., Heylen, D., & Morency, L.P. (2010). Learning and evaluating response prediction models using parallel listener consensus. *Proceeding of International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI)* (p. A3). Beijing, China.
9. de Kok, I., Poppe, R., & Heylen, D. (2012) *Iterative perceptual learning for social behavior synthesis*. Technical Report TR-CTIT-12-01, Enschede.
10. Dittmann, A. T., & Llewellyn, L. G. (1967). The phonemic clause as a unit of speech decoding. *Journal of Personality and Social Psychology*, *6*(3), 341–349.
11. Dittmann, A. T., & Llewellyn, L. G. (1968). Relationship between vocalizations and head nods as listener responses. *Journal of Personality and Social Psychology*, *9*(1), 79–84.
12. Duncan, S, Jr. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*(2), 283–292.
13. Duncan, S, Jr. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, *3*(2), 161–180.

14. Granström, B., House, D., & Swerts, M. (2002). Multimodal feedback cues in human-machine interactions. *Proceedings of the International Conference on Speech Prosody* (pp. 11–14). Aix-en-Provence, France.
15. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R.J., & Morency, L.P. (2006). Virtual rapport. *Proceedings of the International Conference on Interactive Virtual Agents (IVA)* (pp. 14–27). Marina del Rey, CA.
16. Gravano, A., & Hirschberg, J. (2009). Backchannel-inviting cues in task-oriented dialogue. *Proceedings of Interspeech* (pp. 1019–1022). Brighton, UK.
17. Heylen, D., Bevacqua, E., Pelachaud, C., Poggi, I., Gratch, J., & Schröder, M. (2011). Generating Listening Behaviour (Part 4). In: Emotion-oriented systems cognitive technologies, (pp. 321–347) Springer.
18. Huang, L., Morency, L.P., & Gratch, J. (2010). Learning backchannel prediction model from parasocial consensus sampling: A subjective evaluation. *Proceedings of the International Conference on Interactive Virtual Agents (IVA)* (pp. 159–172). Philadelphia, PA.
19. Huang, L., Morency, L.P., & Gratch, J. (2011). Virtual rapport 2.0. *Proceedings of the International Conference on Interactive Virtual Agents (IVA)* (pp. 68–79). Reykjavik, Iceland.
20. Jonsdottir, G.R., Gratch, J., Fast, E., & Thórisson, K.R. (2007). Fluid semantic back-channel feedback in dialogue: Challenges and progress. *Proceedings of the International Conference on Interactive Virtual Agents (IVA)* (pp. 154–160). Paris, France.
21. Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, *26*(1), 22–63.
22. Kitaoka, N., Takeuchi, M., Nishimura, R., & Nakagawa, S. (2005). Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems. *Transactions of the Japanese Society for Artificial Intelligence*, *20*(3), 220–228.
23. Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech*, *41*(3–4), 295–321.
24. Maatman, M., Gratch, J., & Marsella, S. (2005). Natural behavior of a listening agent. *Proceedings of the International Conference on Interactive Virtual Agents (IVA)* (pp. 25–36). Kos, Greece.
25. Morency, L. P., de Kok, I., & Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, *20*(1), 80–84.
26. Noguchi, H., & Den, Y. (1998). Prosody-based detection of the context of backchannel responses. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)* (pp. 487–490). Sydney, Australia.
27. Okato, Y., Kato, K., Yamamoto, M., & Itahashi, S. (1996) Insertion of interjectory response based on prosodic information. *Proceedings of the IEEE Workshop Interactive Voice Technology for Telecommunication Applications* (pp. 85–88). Basking Ridge, NJ.
28. Poppe, R., ter Maat, M., & Heylen, D. (2012). Online behavior evaluation with the switching wizard of oz. *Proceedings of the International Conference on Interactive Virtual Agents (IVA)* (pp. 486–488). Santa Cruz, CA.
29. Poppe, R., Truong, K.P., & Heylen, D. (2011). Backchannels: Quantity, type and timing matters. *Proceedings of the International Conference on Interactive Virtual Agents (IVA)* (pp. 228–239). Reykjavik, Iceland.
30. Poppe, R., Truong, K.P., Reidsma, D., & Heylen, D. (2010). Backchannel strategies for artificial listeners. *Proceedings of the International Conference on Interactive Virtual Agents (IVA)* (pp. 146–158). Philadelphia, PA
31. Truong, K.P., Poppe, R., & Heylen, D. (2010). A rule-based backchannel prediction model using pitch and pause information. *Proceedings of Interspeech* (pp. 490–493). Makuhari, Japan.
32. Truong, K.P., Poppe, R., de Kok, I., & Heylen, D. (2011). A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. *Proceedings of Interspeech* (pp. 2973–2976). Florence, Italy.
33. Valstar, M.F., McKeown, G., Cowie, R., & Pantic, M. (2010). The Semaine corpus of emotionally coloured character interactions. *Proceedings of the International Conference on Multimedia & Expo* (pp. 1079–1084). Singapore, Singapore.
34. Van Welbergen, H., Reidsma, D., Ruttkay, Z., & Zwiers, J. (2010). Elckerlyc: A BML realizer for continuous, multimodal interaction with a virtual human. *Journal of Multimodal User Interfaces*, *3*(4), 271–284.
35. Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, *32*(8), 1177–1207.
36. Xudong, D. (2009). Listener response. In: The pragmatics of interaction (pp. 104–124). Amsterdam: John Benjamins Publishing.
37. Yngve, V.H. (1970). On getting a word in edgewise. In: Papers from the Sixth Regional Meeting of Chicago Linguistic Society, pp. 567–577. Chicago Linguistic Society.