Timing and Entrainment of Multimodal Backchanneling Behavior for an Embodied Conversational Agent

Benjamin Inden
Artificial Intelligence Group, Bielefeld University
Universitätsstr. 25, 33615 Bielefeld, Germany
binden@techfak.uni-bielefeld.de

Zofia Malisz
Petra Wagner
Faculty of Linguistics and Literary Studies, Bielefeld University
{zofia.malisz, petra.wagner}@uni-bielefeld.de

Ipke Wachsmuth
Artificial Intelligence Group, Bielefeld University
ipke@techfak.uni-bielefeld.de

ABSTRACT

We report on an analysis of feedback behavior in an Active Listening Corpus as produced verbally, visually (head movement) and bimodally. The behavior is modeled in an embodied conversational agent and displayed in a conversation with a real human to human participants for perceptual evaluation. Five strategies for the timing of backchannels are compared: copying the timing of the original human listener, producing backchannels at randomly selected times, producing backchannels according to high level timing distributions relative to the interlocutor's utterance and pauses, or according to local entrainment to the interlocutors' vowels, or according to both. Human observers judge that models with global timing distributions miss less opportunities for backchanneling than random timing.

Categories and Subject Descriptors

H.5.2 [**Information Systems**]: Information Interfaces and Presentation—*User Interfaces*

Keywords

embodied conversational agents; backchannels; entrainment

1. INTRODUCTION

1.1 Backchanneling for embodied conversational agents

Embodied conversational agents (ECAs) offer great perspectives for improving human-computer interaction in various tasks including information desk systems, personal assistants and cooperative construction [21]. However, to maximize productivity and long time stability of spoken interaction, attention has to be paid to the naturalness of communicative behavior displayed by the ECA. This includes many aspects, among them the ability of the ECA to generate active listener behavior and produce proper feedback signals [20, 23]. Verbal and visual feedback signals are paramount to establishing rapport and updating information structure (grounding). Those feedback signals that encourage the dialogue partner to continue speaking and smoothly, swiftly and continually yield the turn are called backchannels. The present work focuses on modeling backchanneling.

ICMI'13, December 9–12, 2013, Sydney, Australia

ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2522890.

Truong et al. [31] have studied vocal, visual and bimodal backchannels in their prosodic context as well as gaze pattern effects on the presence and absence of listener feedback. Since head nods are not interfering with the interlocutor's speech, one would expect that they are placed throughout the discourse [31]. However, in the corpus material of spontaneous conversations in Dutch studied by Truong et al. [31] they found that regardless of modality the probability of feedback becomes simply higher as the interlocutor utterance progresses. The presence of visual feedback is closely linked with mutual gaze thus correlating with active listening displays. Truong et al. [31] found that in face-to-face communication the effect of pitch contours in cueing backchannels is less significant than mutual gaze. Especially visual backchannels (head gestures) were significantly more often timed with mutual gaze. Their results also showed the tendency for vocal backchannels to occur during interlocutor pauses rather than interlocutor turns.

Several backchanneling strategies that depend on the speaker's pitch, pauses in the speaker's speach, and gaze interaction, were evaluated on an ECA in a previous study [28]. In that study and a number of followup studies, it was found that a strategy that just copies the timings of the original listener is often perceived as more natural than a strategy based on hand-designed rules [26, 27]. The studies also suggest that random backcanneling according to an Erlang distribution achieves a rather good perceptual naturalness rating from human observers. The quantity of backchannels was also found to be a significant factor influencing perceived naturalness.

In another approach, backchannel behavior from several individuals was collected to provide a backchannel signal for an ECA by means of the so-called *parasocial consensus sampling* [15]. Subsequently, data obtained from parasocial consensus sampling was used to train a machine learner. The features used for training were established in a previous experiment using automatic feature selection. Backchannel-

ing behavior produced after training received better ratings from human subjects than a random strategy or one generating backchannels according to a hand-designed rule [14, 24].

There is also a report on using so-called *iterative perceptual learning* for appropriate timing of backchannels. This method iterates between phases of learning backchannel opportunities from positive and negative examples by a standard machine learning technique. It harvests more positive and negative examples by displaying the learned behavior to human subjects and recording their evaluations on the appropriateness of individual backchannels [11].

1.2 Backchannel timing and entrainment

The present work aims to complement earlier approaches on generating backchannel timings by taking into account the hypothesis that backchannel timings, like the timings of many other events in natural dialogues, are strongly influenced by mutual entrainment. By entrainment, we mean the adaptation of phases and periods of oscillatory movements between the speaker and the listener [32, 33]. Entrainment is a mechanism of coordination and often emerges as a property of oscillating systems. Entrainment phenomena have been detected on various levels of inter-speaker coordination, e.g. in synchronous speech reaching very high temporal agreement [9], in the timing of overlapping speech [37] or in postural swing [29], but also for various other group interactions such as spontaneous rhythmic clapping [25].

Similarly, there have been many studies postulating that listeners use their dialogue partners' speech rhythm and time their feedback responses or swift turn taking along with the beat [8]. However they did not receive empirical support, perhaps because they assumed a high level of strict periodicity in the partner's speech, and used averaging over intervals as basis for prediction of the feedback response timing [6, 2].

The entrainment approach does not necessitate interpolating from the rhythm of the dialogue partner in such a way but uses the dynamics of prosodic events in speech as they unfold to adapt its timing predictions. Along these lines, an ECA has already been presented that moves according to the rhythm of music as processed by entraining oscillators [17]. Models that involve entrainment are useful in that they can be helpful in the anticipation of turn ends or temporal windows, for example for backchannels, while dynamically adapting to speech tempo by period adaptation [16]. It has been proposed that endogenous oscillators in the brains of the speaker and the listener become mutually entrained on the basis of the speaker's rate of syllable production governing readiness for taking the turn at any given instant ([34], see also recent results by [19]). According to that particular model, readiness functions of the listener are counterphased with that of the speaker, and entrainment continues briefly after speech ceases. The hypotheses included in [34] have not been extensively tested via a modeling approach. This work aims at filling this gap by providing prosodic event timings produced by the dialogue partner as input to an abstract oscillator that in turn provides a timing prediction for a listener reponse in the immediately following pause (see section 3 for details). In summary, formal models of oscillator entrainment provide a good explanatory basis as well as testable hypotheses constraining the temporal co-ordination between speaker and listener.

In this work, we first analyse a corpus of Active Listening in order to generate models of general timing distributions of backchannels with respect to the interlocutor's phrases. Regularities with respect to these high level rhythmic structures can be considered as providing evidence of high level entrainment to the interlocutor (with backchanneling being phase-shifted as in the above discussed previous model [34]). We consider several backchanneling modalities: visual (head movement), verbal and bimodal. We complement these high level timing distributionss with local timing predictions on a smaller scale by entraining to the vocalic onsets of the interlocutor the listener is interacting with. We implement these global and local timing strategies in an artificial agent and evaluate the rapport, attention, timing accuracy and missed backchannel opportunities (following [15]) in a study with human participants.

2. EMPIRICAL DATA ON BACKCHANNELING

2.1 The Active Listening Corpus

Nine dialogues from the Active Listening Corpus (ALiCo) were analysed to find global timing patterns of feedback responses in three modalities: verbal, visual (head movement) and bimodal relative to the speaker's utterances and pauses. The analysis was carried out on a German corpus of face-to-face conversations described in detail in [7]. The randomly assigned dialogue partners were given two different roles: the storyteller told two holiday stories to the dialogue partner, the listener, who was instructed to listen actively and participate in the dialogue. The corpus was collected for the purposes of modeling entrainment in dialogue, multimodal behavior of the listener (henceforth the Listener), i.e., feedback signals, head and manual gesture, as well as the prosody of the storyteller (henceforth the Speaker).

Audiovisual recordings were made in a sound-treated studio. Participants were positioned approximately three meters apart to minimize crosstalk. Close talking high-quality headset microphones were used.

All speech annotations were performed in Praat [4] independently from the head gesture annotations. Short spoken feedback expressions produced by the listener and the corresponding feedback function labels were extracted unchanged from annotations described in [7]. Listener contributions not marked as feedback (listener turns) were excluded from the present analysis.

The corpus is equipped with annotations of the Speaker's utterances and pauses labeled manually according to the Rhythm and Pitch (RaP) system of prosodic annotation [5]. An utterance boundary was placed each time a minimally perceptible disjuncture in the flow of speech was determined (i.e., not at all acoustic pauses). The phrases delimited this way approximately correspond to minor intonation phrases in ToBI systems, according to [5]. The minimum pause duration of 50msec resulting from our RaP based annotation is comparable to automatic annotation of Interpausal Units (IPUs) as used in e.g. [3].

Careful annotation of the acoustic signal makes it possible to approximate emergent rhythmic phenomena [12]. To represent the syllabic oscillator hypothesized for speech production, vowel onsets were extracted semi-automatically from the data [10, 1] and checked for accuracy. Next, experts

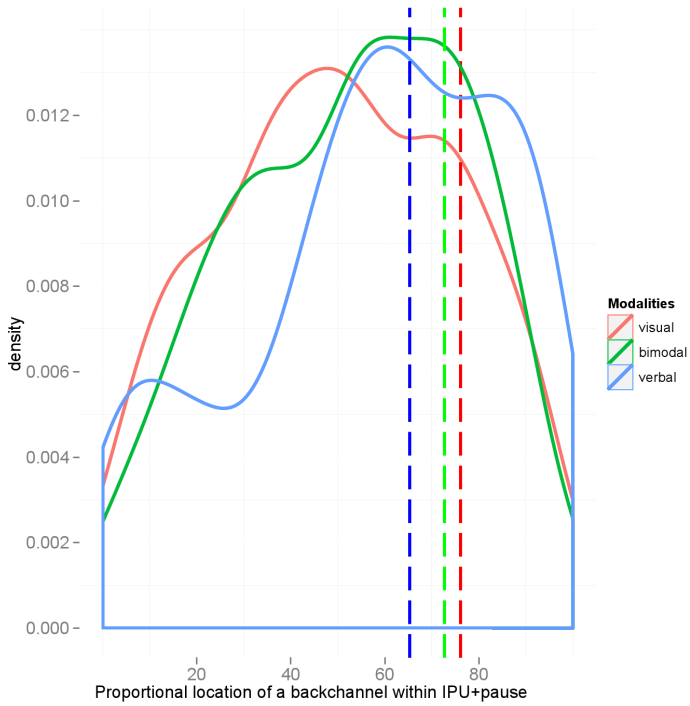

Figure 1: Distribution of backchannels within speaker utterances and subsequent pauses. The end of the phrase is marked by a vertical dotted line.

annotated rhythmic prominence intervals, representing the slower stress oscillator, where each prominent syllable is a pulse on that level. The annotation was based on perceptual judgments of the signal, i.e.: a prominent syllable was marked when a "beat" on a given syllable was actually perceived and not when phonological rules dictated lexical or sentence stress placement. The prosodic annotations of the speaker served as input to the modeling of local timing for the ECA.

Finally, the onsets of head nods and verbal backchannels were also annotated. Annotation of head movement behavior was performed in ELAN (`http://tla.mpi.nl/tools/tla-tools/elan/`, see [35]) by close inspection of the muted video. Perceptually coherent and continuous movements, i.e. Head Gesture Units (HGUs) were segmented first (see [36] for details). The structure of an HGU consists of movement type and the number of movement cycles e.g. nod-2+jerk-1. Four annotators segmented and labeled the listener HGUs independently. Each completed annotation was checked for errors by two other annotators in rotation but no inter-annotator agreement was calculated. However, an evaluation of the head annotation scheme and inter-annotator consistency for a different spontaneous dialogue dataset can be found in [22]. An inter-annotator agreement for HGUs identification was found to be 77%. Duration agreement yielded 79% in [22] indicating consistency among annotators in marking gesture boundaries and event identification with this scheme.

2.2 Analysis and results

First, we excluded all listener HGUs that overlapped with listener turns (18.3%) to ensure we process feedback signals and not turn-taking. In total, 1578 head gesture units were identified (Speaker N= 1001, Listener N=577) in 9 dialogues. Verbal feedback expressions produced by the listener equaled 514 and the total number of speaker's utterances was 1049. The dialogues have a total length of 66 minutes with a mean length of 7:30 min. (Min = 6:00, Max=8:50, SD=1:05). On average, 15 HGUs per minute were produced by the speaker. The listener's feedback rate was 9 per minute for head movements (bimodals included) and 8 per minute for purely verbal responses.

We assume that in this type of dialogue the minimal unit of analysis within which grounding is aimed to be achieved is the storyteller's utterance and the following pause. We studied feedback events relative to these two subsequent units. Single overlaps between visual feedback and verbal feedback that overlapped completely with either speaker's utterance or pause were included in the bimodal set. The visual set contained head gestures whose onset fell into the unit under consideration, i.e., if the gesture spanned several units, only the first unit overlapping the gesture onset was included.

Forty percent of head gestures in the visual feedback data start within the utterance but also overlap the following pause. The median proportional position within the Speaker phrase for these head gestures is 85%. We believe the skew in the timing distribution of head gestures towards the end of the Speaker's phrase is caused by the doubly overlapping gestures. We suggest that many head gestures that begin at the very final portion of Speaker utterances are located already at a position where potential turn-taking negotiation is taking place between the interlocutors. The visual responses taking place well within the turn function as continuers. We therefore look at the distribution of only those visuals that completely fall into the phrase.

Recall that Truong et al. [31] concluded that regardless of modality the probability of feedback is simply higher as the interlocutor utterance progresses, also in case of head nods. Figure 1 presents the distribution of the three modalities in our Active Listening Corpus within an utterance and the subsequent pause. What is clear is that in general, the majority of listener feedback indeed rises towards and coincides with the ends of phrases and beginning of pauses, especially feedback signals with a spoken component (verbals and bimodals). The head gesture data however is more uniformly distributed when visual and bimodal feedback that does not co-overlap the subsequent pause is included. The result supports the notion that some head gestures are placed throughout the Speaker's utterance when functioning as an unobtrusive visual backchannel.

We subsequently utilize the resulting timing distributions for the three modalities as timing models for an implementation in an ECA.

3. MODELING METHODS

3.1 Basic approach

From the results of the corpus analysis, the global backchanneling strategy displayed in Fig. 2 was designed: Visual backchannels are broadly distributed within the speaker's utterances, wheras verbal backchannels are normally distributed around the end of utterances. For the purpose of

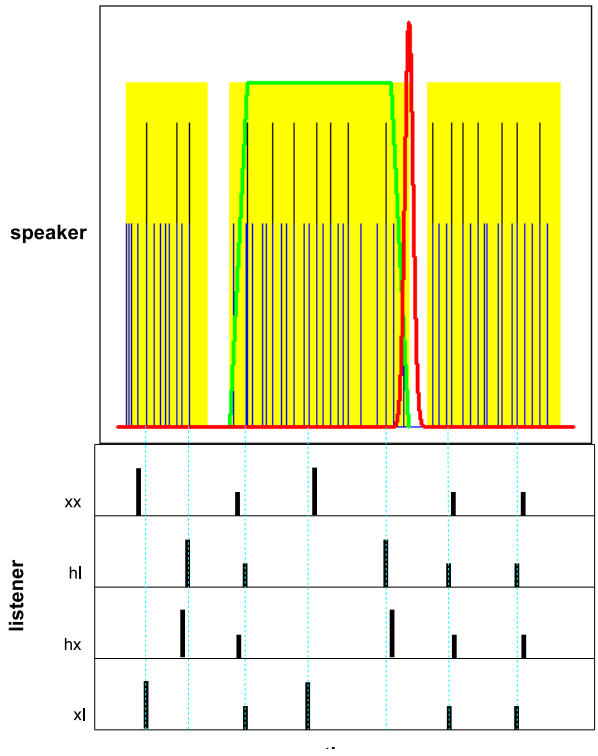

Figure 2: A simplified approach towards rhythmic timing of backchannels. The upper panel shows a part of a dialog with three phrases. Phrases are displayed in yellow, vowel onsets in blue (lower spikes), feet in black (higher spikes), the visual backchannel distribution in green (trapezoid curve), and the verbal backchannel distribution in red (Gaussian curve). The lower panel shows schematic examples of four of the backchanneling strategies explored here: xx, random; hl, high level timing and local entrainment; hx, high level timing only; xl, local entrainment only. Vocal backchannels are displayed as long bars, whereas visual backchannels are displayed as short bars. The dotted lines indicate the exact timings of rhythmic events derived from the speaker. Note that strategies involving local entrainment move backchannels to these exact timings as opposed to the random strategy. Strategies involving high level distributions tend to move vocal backchannels to the ends of the speaker's phrases.

this experiment, bimodal utterances are treated just like verbal backchannels. We are well aware that this is a strongly simplified and idealized picture of BC timings, which can also be positioned differently with respect to the various rhythmic units for reasons of dialogue content or social interactions not captured by our approach. Nevertheless, here we explore whether these empirically derived rules of rhythmic timing can already achieve a significant improvement in terms of perceived naturalness of ECA backchanneling behavior.

3.2 Backchannel Generation strategies

3.2.1 Random timing (xx strategy)

A sequence of BC timings is created by starting at the end of the first IPU and creating a BC every 6.0 seconds until the end of the dialogue is reached. The timings are then perturbed randomly with a uniform distribution of ± 2.0 seconds each, and a type is determined randomly according to the following probabilities that approximate data from the corpus: head nod, 0.55; verbal BC, 0.35; bimodal BC, 0.1. This timing method serves to ensure an approximately even backchanneling behavior over coarse time scales.

In addition to the described backchannels, eye blinks are also part of the ECA behavior. Their timings are generated by starting at the second vowel onset and creating a prospective event every 5.0 seconds until the end of the dialogue is reached. The timings are then perturbed randomly with a uniform distribution of ± 1.0 seconds each.

To make the behavior of the ECA look more natural, the verbal utterance 'um' was randomly chosen from among two versions with slightly different prosody each time it was generated.

3.2.2 High level timing / local timing with entrainment (hl strategy)

The random timings from section 3.2.1 are taken as input, and the IPU whose center is closest to the prospective BC timing is identified. Informed by results of the corpus analysis, it is assumed that the onset times of verbal and bimodal BCs are normally distributed around the end of an IPU, with a standard deviation of 0.1 seconds. The visual backchannels, on the other hand, are assumed to be distributed according to a distribution that plateaus in the inner part of a given IPU.

Next, the local timing of a given BC in this strategy is determined as follows: for visual backchannels, all vowel onsets of the speaker within the given IPU are considered candidates for a BC onset. One of them is chosen, with probabilities being proportional to the height of the visual BC probability distribution at the times of the respective candidates. For verbal and bimodal BCs, all rhythmic prominence onsets of the speaker within 5 standard deviations around the end of the given IPU are considered candidates. One of them is chosen, with probabilities being proportional to the height of the normal distribution at the times of the respective candidates. In the case of a bimodal BC, the visual component is placed on the vowel onset immediately preceeding the rhythmic prominence onset where the verbal component has been placed.

The above description assumes that there are vowel onsets and rhythmic prominences before and after the end of the IPU. But of course, there are none in a pause, so the rhyth-

mic events within pauses are assumed to be derived from an entrained oscillator. The particular technique chosen here works as follows: for generating event timings within a given pause, all rhythmic events from the respective level in the preceding IPU are considered. If there are at least two of them, the pause is filled by an abstract oscillator starting with a phase of 0.0 at the last event within the IPU. The period of the oscillator is calculated from a weighted mean of the intervals between all successive events in the preceding IPU, where the weights decay exponentially with distance from the pause. Specifically, we set the weight of the ith cycle before the pause to 0.9^i, and normalize by dividing by the sum of all weights. In reality, different oscillator models will produce slightly different timings due to various non-linearities, but here we assume that the differences are so small that they can be disregarded if only short pauses are bridged, as we do here.

The eyeblink timings generated by the random strategy are adjusted by choosing the vowel onset of the speaker closest to the preliminary BC.

3.2.3 *High level timing only (hx strategy)*

The timings from the previous section are randomly perturbed by ± 0.05 seconds, which means that they are not aligned with rhythmic events like vowel onsets and rhythmic prominences any more, but are still where they would be expected according to their global distribution within phrases.

3.2.4 *Local timing with entrainment only (xl strategy)*

For this strategy, the timings from the random strategy (section 3.2.1) are adjusted such that they align with the nearest rhythmic event (rhythmic prominences in the case of verbal and bimodal backchannels, and vowel onsets in the case of visual backchannels and eye blinks). High level timing distributions are disregarded.

3.2.5 *Copy timing (co strategy)*

This strategy just uses the timings of head nods and verbal backchannels recorded and annotated from the original listener. As there is no annotation of eye blinks available, these are generated in exactly the same way as described for the high level timing / local timing with entrainment (hl) strategy.

3.3 Evaluation

The evaluation procedure is similar to that of previous work by other authors [15]. Three prerecorded dialogues were selected from the corpus, and for each, five videos were generated according to the backchanneling strategies discussed in the previous section, each showing the original speaker and the ECA (see Fig 3). While the original videos had varying lengths, two clips of approximately 30 seconds each were cut out from each video, and used for evaluation. Any sequences where the original listener did anything beyond providing simple backchannels were excluded from selection (sometimes, they took turns by asking a question or completing the speakers' sentences).

Each of the 37 participants saw all five BC strategies for one conversation clip presented in random order, then the next five, until all 30 clips were shown. The order of the groups of five clips belonging to the same original clips was also randomized. The participants were told in advance that they were to evaluate the timings of the head nods and 'um' utterances in each video, and that this was the only difference between the videos. After watching a video, the following questions had to be answered on a scale between 1 and 7:

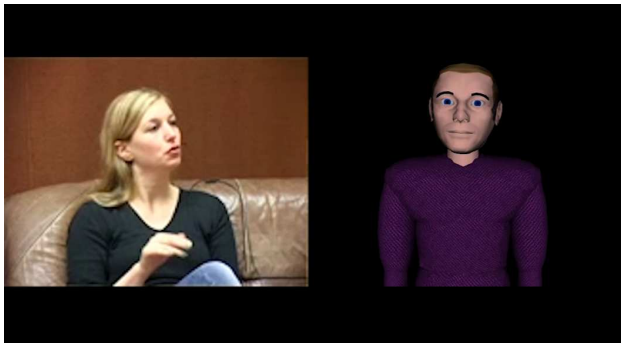

Figure 3: The conversation scenario used for evaluation.

- How much rapport did you feel between the ECA and the speaker while watching the video?
- Do you believe the ECA was listening carefully to the speaker?
- How often do you think the ECA nodded or said 'um' at inappropriate times?
- How often do you think the ECA missed opportunities for nodding or saying 'um'?

After watching all the videos, participants also had to indicate how difficult they found the task on a scale from 1 to 7.

4. RESULTS

We examined which backchanneling strategies were rated better with respect to the four questions as compared to the random timing strategy as a baseline. To achieve this goal, a linear mixed model was fitted to the data for each question using the statistics package R and the R library lme4. These linear mixed models take the following additional factors into account: the order of presentation; the ID of the original conversation video clip (six different values); and the sum of all backchannels in a given synthesized clip (the copy strategy often produced much more backchannels than the other strategies, while the number of backchannels generated by the other strategies could vary slightly due to small timing differences at the beginning or end of a clip). We also entered the evaluating participants as the random effect into all models. [1]

[1]Random effects residual variance for perceived rapport is 1.5, (standard deviation = 1.23); for perceived attentiveness, variance = 1.88 (sd = 1.4); for perceived wrong placements, variance = 2.24 (sd = 1.5); and variance = 1.64 (sd = 1.3) for perceived missed opportunities.

4.1 Overall task difficulty

Participants reported a task difficulty of 3.8 on average (sdev=1.4). A fair number reported that they found it difficult to concentrate on so many and so mutually similar videos. There were also some complaints about the lack of smiling in the ECA — something that we do not model here, but which certainly would increase the perceived naturalness.

4.2 Perceived rapport

The model fitted for rapport also takes gender of the participant into account because exploration of other models had shown that this is a significant factor for this question. The model produced the following estimates for factors that significantly improve perceived rapport as compared to the baseline: presentation order (effect size $b = 0.01$, $p = 0.009$), the copy strategy ($b = 0.62$, $p < 0.001$), female gender ($b = 0.66$, $p = 0.011$). An increasing sum of backchannels decreased perceived rapport ($b = -0.04$, $p = 0.011$). Many of the video clips were also rated as significantly different from the baseline clip. However, the timing strategies did not differ significantly from the random strategy in terms of perceived rapport.

4.3 Perceived attentiveness

The linear mixed model finds that the copy strategy significantly improves perceived performance as compared to the random backchanneling strategy ($b = 0.67$, $p < 0.001$), whereas an increasing sum of backchannels slightly decreases the rating ($b = -0.04$, $p = 0.014$). No other backchanneling strategy differs significantly from the random strategy. Again, some particular video clips got significantly different ratings.

4.4 Perceived wrong placements

The only significant simple factor here is the sum of backchannels, an increasing number of which also increases perceived wrong timings ($b = 0.51$, $p < 0.001$). While the values for the copy strategy are not significant ($b = 0.69$, $p = 0.25$), there is a significant influence of the interaction of this strategy and the number of backchannels ($b = -0.32$, $p = 0.010$). All other strategies do not achieve a significant change when considered alone or in combination with the number of backchannels. Again, some of the different clips got significantly different ratings.

4.5 Perceived missed opportunities

Here an increasing number of backchannels significantly decreases perceived missed opportunities ($b = -0.69$, $p < 0.001$). The copy strategy ($b = -3.30$, $p < 0.001$), the high/local timing (hl) strategy with entrainment ($b = -1.73$, $p < 0.001$), and the high level timing only (hx) strategy ($b = -1.50$, $p = 0.001$) significantly improve performance, while the local timing only (xl) strategy does not perform significantly differently from the random strategy ($b = -0.79$, $p = 0.100$). There are also significant negative influences of the number of backchannels in interaction with the copy ($b = 0.56$, $p < 0.001$), high/low level (hl) ($b = 0.40$, $p < 0.001$) and high level only (hx) ($b = 0.34$, $p < 0.001$) strategies. This indicates that as the number of backchannels increases, the advantages of those strategies over the random strategies decreases.

5. DISCUSSION

We have shown that those strategies using empirically derived global timing distributions are perceived as missing less opportunities for backchanneling than a random strategy. Effect sizes suggest that a strategy that combines high level and local timing with entrainment improves the ratings of missing less feedback opportunites more strongly than the one based on high level timing only. We could not demonstrate a positive effect of local timing alone over random timing.

In line with previous studies, we find that the number of backchannels influences all aspects of perceived backchanneling behavior strongly and significantly.

We also do not find a positive effect of high level timing on other aspects of perceived backchanneling behavior such as perceived attentiveness or rapport between the ECA and the speaker. The copy strategy, on the other hand, does perform significantly better on all aspects than the random strategy.

It is posssible that advantages of low level entrainment of backchannels to the speaker's vowels could not be detected because there were to many different strategies presented for evaluation. Also, the variation of the sum of backchannels across different video clips might have influenced the participants' ability to discern subtle timing effects. Therefore, we plan to proceed by doing a simpler and more controlled perceptual evaluation of local, low level timing with entrainment in the future.

The timings for backchanneling have been generated offline for this study. This is sufficient for studying user preferences, but the ultimate goal is to produce backchannels online. The strategies introduced here can be generated online if two conditions are met. Firstly, speaker rhythms have to be extracted online. Online extraction of vowel onsets has been done before [38, 13]. Automatic prominence detection can also be done online [30]. Secondly, rhythmic evens in the future have to be predicted. In the context of the strategies introduced here, this is necessary for predicting the length of the phrase as well as for generating rhythmic timings in the speaker's pauses. Oscillators that entrain to the speaker's rhythm can be used for making such predictions [16].

The strategies introduced here do not perform as well as a human strategy copied by the ECA. This is not surprising given that the human can evaluate various semantic and gestural cues that are not available in our framework. As stated in the introduction, the strategies explored here are complementary to some previous approaches, which means they could be combined to further enhance the performance. For example, instead of generating bachchannels with approximately regular spacing on coarse time scales as done here, the change in a speaker's pitch, eye contact to the listener, or keyword spotting could modify the probabilities of BC generation [31, 18].

Acknowledgments

This research is kindly supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673.

6. REFERENCES

[1] P. A. Barbosa. *Incursões em torno do ritmo da fala.* Pontes, Campinas, 2006.

[2] S. Benus. Adaptation in turn-initiation. In A. E. et al., editor, *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues.*, Lecture Notes in Computer Science [LNCS 6456], pages 72–80. Springer Verlag, Berlin, Heidelberg, 2011.

[3] S. Benus, A. Gravano, and J. Hirschberg. The prosody of backchannels in American English. In *Proceedings of the 16th International Congress of Phonetic Sciences*, pages 1065–1068, Saarbrücken, Germany, 2007.

[4] P. Boersma and D. Weenink. Praat: Doing phonetics by computer. version 5.3.04, 2012.

[5] M. Breen, L. Dilley, J. Kraemer, and E. Gibson. Inter-transcriber reliability for two systems of prosodic annotation: Tobi (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory*, 8:277–312, 2010.

[6] M. Bull and M. Aylett. An analysis of the timing of turn-taking in a corpus of goal-orientated dialogue. In *Proceedings of ICSLP*, pages 1775–1778, Sydney, 1998.

[7] H. Buschmeier, Z. Malisz, M. Włodarczak, S. Kopp, and P. Wagner. 'Are you sure you're paying attention?' – 'Uh-huh'. Communicating understanding as a marker of attentiveness. In *Proceedings of Interspeech 2011*, pages 2057–2060, Florence, Italy, 2011.

[8] E. Couper-Kuhlen. *English Speech Rhythm: Form and function in everyday verbal interaction.* Benjamins, Amsterdam, 1993.

[9] F. Cummins. Practice and performance in speech produced synchronously. *Journal of Phonetics*, 31:139–148, 2003.

[10] F. Cummins and R. Port. Rhythmic constraints on stress timing in english. *Journal of Phonetics*, 26:145–171, 1998.

[11] I. de Kok, R. Poppe, and D. Heylen. Iterative perceptual learning for social behavior synthesis. 2013.

[12] D. Gibbon and F. R. Fernandes. Annotation-mining for rhythm model comparison in brazilian portuguese. In *Proceedings of INTERSPEECH*, 2005.

[13] C. Heinrich and F. Schiel. Estimating speaking rate by means of rhythmicity parameters. In *Proceedings of Interspeech*, 2011.

[14] L. Huang, L.-P. Morency, and J. Gratch. Learning backchannel prediction model from parasocial consensus sampling: A subjective evaluation. In *Proceedings of the International Conference on Intelligent Virtual Agents*, 2010.

[15] L. Huang, L.-P. Morency, and J. Gratch. Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. In *Proceeedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, 2010.

[16] B. Inden, Z. Malisz, P. Wagner, and I. Wachsmuth. Rapid entrainment to spontaneous speech: A comparison of oscillator models. In *Proceedings of the Cognitive Science Conference*, 2012.

[17] I. Jauk, P. Wagner, and I. Wachsmuth. Dynamic perception-production oscillation model in human-machine communication. In *Proceedings of the International Conference on Multimodal Interaction*, 2011.

[18] G. Jonsdottir, J. Gratch, E. Fast, and K. Thórisson. Fluid semantic back-channel feedback in dialogue: Challenges and progress. In *Proceedings of the International Conference on Interactive Virtual Agents*, 2007.

[19] M. Kawasaki, Y. Yamada, Y. Ushiku, E. Miyauchi, and Y. Yamaguchi. Inter-brain synchronization during coordination of speech rhythm in human-to-human social interaction. *Nature*, 3, 2013.

[20] S. Kopp, J. Allwood, K. Grammer, E. Ahlsen, and T. Stockmeier. Modeling embodied feedback with virtual humans. In I. Wachsmuth and G. Knoblich, editors, *Modeling communication with robots and virtual humans*. Springer-Verlag Berlin Heidelberg, 2008.

[21] S. Kopp, B. Jung, N. Leßmann, and I. Wachsmuth. Max — a multimodal assistant in virtual reality construction. *Künstliche Intelligenz*, 4:11–17, 2003.

[22] S. Kousidis, Z. Malisz, P. Wagner, and D. Schlangen. Exploring annotation of head gesture forms in spontaneous human interaction. In *TiGeR 2013, Tilburg Gesture Research Meeting*, 2013 (submitted).

[23] R. M. Maatman, J. Gratch, and S. Marsella. Natural behavior of a listening agent. In *Proceedings of the International Conference on Interactive Virtual Agents*, pages 25–36, 2005.

[24] L.-P. Morency, I. de Kok, and J. Gratch. Predicting listener backchannels: A probabilistic multimodal approach. In *Proceedings of the 8th International Conference on Intelligent Virtual Agents*, 2008.

[25] Z. Néda, E. Ravasz, T. Vicsek, Y. Brechet, and A. L. Barabasi. The physics of rhythmic applause. *Physical Review E*, 61:6987, 2000.

[26] R. Poppe, K. P. Truong, and D. Heylen. Bbackchannel: Quantity, type and timing matters. In *Proccedings of Intelligent Virtual Agents*, 2011.

[27] R. Poppe, K. P. Truong, and D. Heylen. Perceptual evaluation of backchannel strategies for artificial listeners. *Autonomous Agents and Multi-Agent Systems*, 2013.

[28] R. Poppe, K. P. Truong, D. Reidsma, and D. Heylen. Backchannel strategies for artificial listeners. In *Proceedings of the Intelligent Virtual Agents Conference*, 2010.

[29] C. Richardson, R. Dale, and K. Schockley. Synchrony and swing in conversation: Coordination, temporal dynamics, and communication. In I. Wachsmuth, M. Lenzen, and G. Knoblich, editors, *Embodied Communication in Humans and Machines*. Oxford University Press, 2007.

[30] F. Tamburini and P. Wagner. On automatic prominence detection for german. In *Proceedings of INTERSPEECH*, 2007.

[31] K. P. Truong, R. Poppe, I. de Kok, and D. Heylen. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In *Proceedings of INTERSPEECH*, 2011.

[32] I. Wachsmuth. Communicative rhythms in gesture and speech. In *Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, 1999.

[33] P. Wagner, Z. Malisz, B. Inden, and I. Wachsmuth. Interaction phonology — a temporal co-ordination component enabling communicative alignment. In *Towards a New Theory of Communication*. John Benjamins, to appear.

[34] M. Wilson and T. P. Wilson. An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review*, 12:957–968, 2005.

[35] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. Elan: A professional framework for mulmultimodal research. In *Proceedings of the fifth International Conference on Language Resources and Evaluation*, 2006.

[36] M. Włodarczak, H. Buschmeier, Z. Malisz, S. Kopp, and P. Wagner. Listener head gestures and verbal feedback expressions in a distraction task. In *Proc. of the Interdisciplinary Workshop on Feedback Behaviours in Dialogue*, 2012.

[37] M. Włodarczak, J. Simko, and P. Wagner. Temporal entrainment in overlapped speech. In *Proceedings of Interspeech*, 2012.

[38] Y. Zhang and J. Glass. Speech rhythm guided syllable nuclei detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.