The Face Speaks: Contextual and Temporal Sensitivity To Backchannel Responses.

Andrew J. Aubrey[†], Douglas W. Cunningham[*], David Marshall[†],
Paul L. Rosin[†], AhYoung Shin[‡] and Christian Wallraven[‡]

[†]School of Computer Science and Informatics, Cardiff University, Cardiff, UK
[*]Brandenburg Technical University Cottbus, Germany
[‡]Korea University, Korea

Abstract. It is often assumed that one person in a conversation is active (the speaker) and the rest passive (the listeners). Conversational analysis has shown, however, that listeners take an active part in the conversation, providing feedback signals that can control conversational flow. The face plays a vital role in these *backchannel responses*. A deeper understanding of facial backchannel signals is crucial for many applications in social signal processing, including automatic modeling and analysis of conversations, or in the development of life-like, effective conversational agents. Here, we present results from two experiments testing the sensitivity to the context and the timing of backchannel responses. We utilised sequences from a newly recorded database of 5-minute, two-person conversations. Experiment 1 tested how well participants would be able to match backchannel sequences to their corresponding speaker sequence. On average, participants performed well above chance. Experiment 2 tested how sensitive participants would be to temporal misalignments of the backchannel sequence. Interestingly, participants were able to estimate the correct temporal alignment for the sequence pairs. Taken together, our results show that human conversational skills are highly tuned both towards context and temporal alignment, showing the need for accurate modeling of conversations in social signal processing.

1 Introduction

The face and head are a crucial aspect of human communication as they contain a wealth of non-verbal cues and are key indicators of emotional state. This has led to a large body of work on facial expression (including head motion) perception and production. Although conversational analysis is traditionally a cognitive science endeavor, there is a growing interest in the automatic recognition and synthesis of conversational behavior, particularly for the creation of virtual conversation agents. Recent reviews, [1], [2], provided a detailed overview of this new research field of social signal processing.

In terms of facial expression research, the majority of work is based on the so-called universal expressions (happiness, sadness, anger, disgust, fear and surprise) defined by Ekman [3]. With the exception of happiness, however, these expressions do not occur with high frequency in *everyday* conversations. In recent years there has been an effort to examine other expressions that occur in conversations with a higher frequency (such as thinking, agreeing, being confused, being bored, *etc.*): [4–8]. Conversational expressions are not limited to movements of facial muscles, they also include global head motion and orientation (*e.g.* to indicate agreement or disagreement) [9] and gaze [10] (*e.g.* to indicate the addressee of a question). Which regions of the face are necessary and sufficient for expression recognition was investigated in [7]. They showed that the motion of different face regions contribute a varying amount to recognition performance. A clear advantage of dynamic over static stimuli for conversational expressions was demonstrated in experiments conducted in [5]. Modeling of conversational expressions therefore needs to take into account the temporal aspects of facial movements.

The term *backchannel* was coined by Yngve [11] and is used to describe the exchange of signals from the listener(s) to the speaker. These signals, which can control conversational flow, are short visual (*e.g.* nod) or vocal (*e.g.* "uh-huh") signals that the listener uses to indicate understanding, disgust, a desire to speak, or interest in the conversation, for example. While backchannel signals can be considered a subset of all feedback signals [12], this work is only concerned with visual backchannel signals (specifically, those of the face, head, and shoulders).

Until recently, studies on backchannel signals have primarily used static facial expression stimuli, such as the work by Baron-Cohen *et al.* [4]. Recently, Wehrle *et al.* [8] compared dynamic and static expressions. Even though the dynamic data were synthesised, results showed that static stimuli were more easily confused than dynamic. Bavelas *et al.* [10] found that periods of mutual gaze increased the likelihood of a backchannel occurring. In [13], the effect of quantity, timing and type of backchannel was investigated. Participants were asked to rate whether the reaction of an artificial listener to a real speaker was human-like. Several interesting results were obtained. Too many or too few backchannels per minute reduced the quality of the listener, furthermore, a lower and upper limit of 6 and 12 per minute respectively was suggested. Nods were often more appropriate than vocal signals and the timing of the backchannel influences how human-like the listener was perceived.

The goal of the present work is to further our understanding of the perceptual sensitivity to backchannel responses in conversations. More specifically, we will present two experiments that aim to test the contextual and temporal sensitivity for processing of facial feedback signals. These experiments are intended to provide important contributions towards full spatio-temporal modeling of the non-verbal facial (and head-related) information channel in conversations, complementing previous research on conversational facial expressions (*e.g.*, [7]). These kinds of models will be indispensable for the creation of virtual agents or artificial listeners that can display human-like listener behaviors [6, 14].

The remainder of the paper is organized as follows: Section 2 describes the database from which we derived the backchannel responses, Sections 3 and 4 discuss the two experiments, and Section 5 provides a brief conclusion.

2 Database

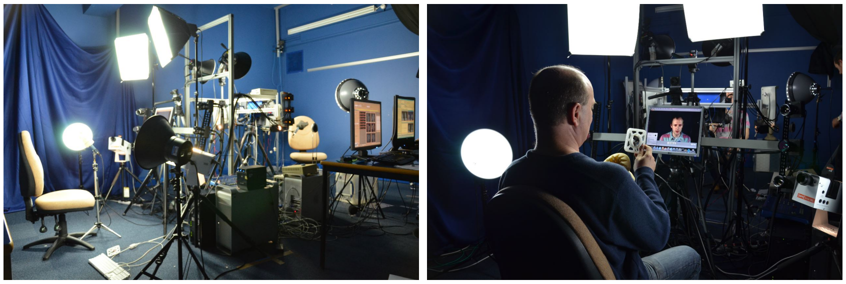

Fig. 1. Left) Setup of recording equipment, Right) View of person opposite during use.

The database contains natural conversations obtained by recording both speaker and listener in a non-scripted conversation.

2.1 Recording Equipment

To capture the conversations in as natural a setting as possible, two audio-video recording systems were set up as shown in Fig. 1(a). The equipment used to capture *each side* of the conversation contained the following: a 3dMD dynamic scanner captured 3D video, a Basler A312fc firewire CCD camera captured 2D color video, and a microphone placed in front of the participant out of view of the camera captured sound (at 44.1KHz). A view from one side of the setup is shown in Fig. 1(b). In this paper only the 2D recordings are used; the 3D system setup and subsequent processing of that data is the subject of future work.

To ensure all audio and video could be reliably synchronized, each speaker had a hand-held buzzer and LED (light emitting diode) device, used to mark the beginning of each recording session. A single button controlled both devices and simultaneously activated the buzzer and LED. No equipment was altered between the recording sessions, except for the height of the chair to ensure the speaker's head was clearly visible by the cameras.

2.2 Recording Methods

The full dataset consists of 30 conversations, each lasting five minutes and containing two people. There were 16 speakers in total, 12 male and 4 female between the ages of 25 and 56. Prior to the recording session each speaker was asked to fill out a questionnaire. The questions simply required a response on a five point scale from strongly dislike (1) to neutral (3) to strongly like (5) and was aimed at finding out how strongly the speakers felt about possible conversation topics. The questionnaire was used to suggest topics to each pair of speakers for which they had similar or dissimilar ratings, and could if they desired be used as a basis for their conversation. Examples of the topics covered in the questionnaire are the like or dislike of different genres of music (rap, opera, jazz, rock etc), literature (poetry, sci-fi, romance biographies etc), movies, art, sports (rugby, football, ice hockey, golf etc), technology (smartphones, tablets), games, television and current affairs. However, the speakers were not restricted to the topics suggested. All participants were fluent in the English language.

2.3 Backchannel Sequences

Eleven short video sequences containing a mixture of speakers and listeners were chosen. Each sequence consisted of a "main channel" (a speaker) and a "backchannel" (a listener's concurrent non-verbal response). The sequences were chosen so that all main channel clips contained a short, easily understandable segment of a conversation. In order to allow us to systematically vary the synchronization between main and backchannel, all backchannel clips were constrained so that they contained one main visible response (possibly followed by several other smaller ones) and no speech for a period of several seconds. This constraint reduced the total number of possible sequences considerably. Sequences 1, 2, 3, 4, 8, 9, and 10 were about movies. Sequences 5 and 6 were about literature. Sequence 7 was about games. Figure 2 (Section 3.3) shows who was involved in each sequence.

3 Experiment 1: Sensitivity to Context

To investigate how well participants can pick the "correct" backchannel response given a main channel spoken segment, five possible main-channel/backchannel pairings were shown to twenty-one participants for each of the eleven sequences. In addition to trying to identify the correct matching, participants also to evaluate each main-channel/backchannel pairing along several dimensions.

3.1 Methods

Stimuli For each of the eleven sequences, we picked four plausible alternate clips using the same listener. Thus, Sequence 1 contained a seven second long snippet of the conversation between S2 (as speaker) and S5 (as listener). The four alternate backchannels also had S5 as a listener. The alternate sequences also contained only one main visible response and no speech for a period of several seconds. The visible response was also to have roughly the same length as the main backchannel, although this was not strictly enforced. This further reduced the possible number of usable backchannel sequences. In some cases the alternate sequences had similar behaviour to the original backchannel (e.g., alternate sequences of agreement or of laughter). In other cases, the alternate sequence was very different, but still a very plausible response. The backchannel of the four alternate sequences were manually synchronized with the main channel. The participants only heard the audio from the main channel.

Procedure The twenty-one participants were seated one at a time in front of a computer screen and asked to wear headphones. The experimental chamber was lit with normal daylight. Once the participant was seated and any initial questions were answered, written instructions for the experiment were then presented. Once participants indicated that they understood, the experiment began (controlled by Psychtoolbox3).

The eleven sequences were shown to the participants in random order, with each participant receiving a different random order. Evaluation of each sequence consisted of three phases, all of which must be completed before another of the eleven sequences could be evaluated. In the first phase, icons representing each of the five main channel-backchannel pairings were shown. Clicking on an icon started a full screen presentation (with the videos of the speaker and the listener being shown side by side) of that particular pairing after which participants were returned to the icons. Participants could watch the pairings as often as they wanted, in any order they wanted. Once all five pairings had been seen at least once each, participants could continue to the second phase.

In the second phase, participants were asked to decide which of the five pairings was the original main-channel/backchannel pairing. In the third phase, participants were shown each of the pairings again, one at a time. After watching the pairing, they were asked to rate it on four different Likert-type scales. The first three scales used the following terms for the five levels "(1) fully inappropriate" , "(2) somewhat inappropriate", "(3) neutral", "(4) somewhat appropriate", "(5) fully appropriate". The first scale asked "How appropriate was the timing of the response?". The second scale asked "How appropriate was the intensity of the response". The third asked "How appropriate was the contents of the response?". The fourth scale asked "How humorous (in terms of the conversational expressions rather than comic dialogue) was the WHOLE conversation?" and used the levels "(1) fully non-humorous", "(2) somewhat non-humorous", "(3) neutral", "(4) somewhat humorous", and "(5) fully humorous".

After the experiment participants were thanked for their participant, paid, debriefed, and any questions were answered.

3.2 Results and Discussion

Overall, people were able to find the correct backchannel and the ratings showed some surprising similarity to the pattern of recognition choices.

3.3 Recognition Performance

With an average performance of 41%, recognition accuracy was significantly above chance; $t(20) = 6.29, p < 0.0001$.

Everyday experience would suggest that some people are more sensitive to the natural flow of a conversation than others. This is reflected in the accuracy results. Most participants were able to correctly identify the original pairing around 40% of the time. Indeed, all but four participants were well above the 20% chance level. Yet, some participants were much more accurate (*e.g.*, participant

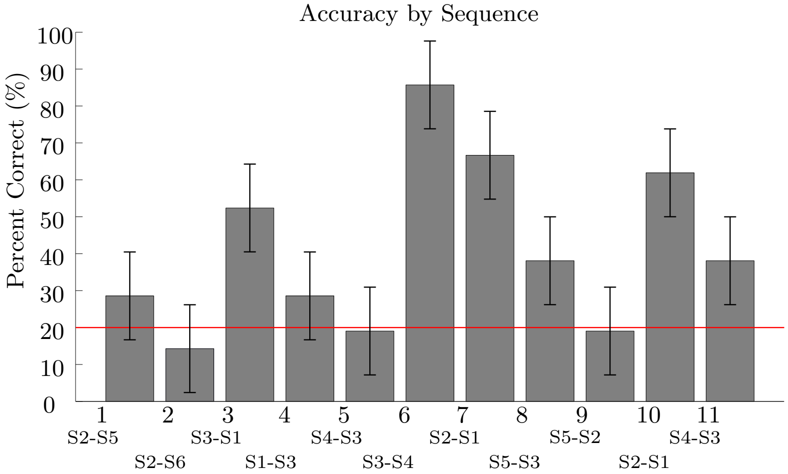

Fig. 2. Accuracy results from Experiment 1, by sequence. The labels indicate the speaker-listener pair. The horizontal line represents chance performance. The error bars represent the 95% confidence interval.

19 at 73%) while others were much worse (*e.g.*, participants 5 and 9, both at 9%).

In Figure 2, the responses are plotted by sequence. There is considerably more variation between the sequences than between the participants. Some sequences (such as Sequence 6) were recognized by almost all participants, while others (Sequence 9) were rarely recognized. Indeed, for at least 5 of the 11 sequences, performance was at or near chance levels! Thus, it seems that the overall recognition rate is being driven by a few exceptional sequences.

The variation between sequences might be due any of a number of reasons, including the degree of talent of the speaker or listener, the topic, or even the poor quality of the alternatives. Eleven sequences is not, however, a large enough sample to conclusively determine why some sequences were better than others. It is clear, however, that some of the recordings of speakers as well as of some listeners were associated with higher high recognition performance. (see Figure 3). This is consistent with everyday experience: some people are better conversationalists than others.

As can be seen in Figure 4 (the leftmost bar in each cluster is always the correct response), some alternatives were more plausible than others. On the other hand, a low quality of the alternatives cannot explain most of the accuracy performance. In most sequences one of the incorrect responses was often chosen, suggesting that these false backchannels did indeed share some characteristics with the proper backchannel. In fact, in 6 of the 11 sequences one false backchannel was chosen more often than the correct alternative.

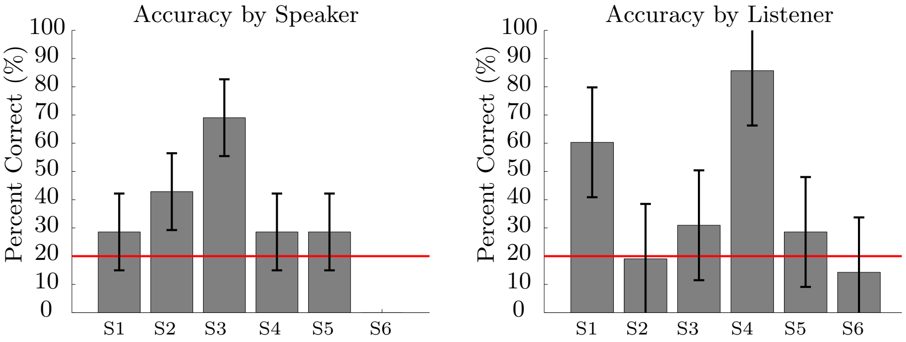

Fig. 3. Accuracy results from Experiment 1, by speaker (Left) and listener (Right). The horizontal line represents chance performance. The error bars represent the 95% confidence interval.

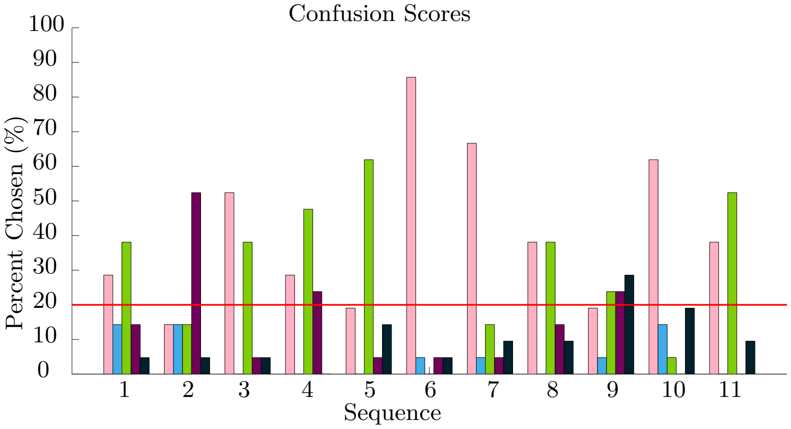

Fig. 4. Confusions in Experiment 1. The horizontal line represents chance performance.

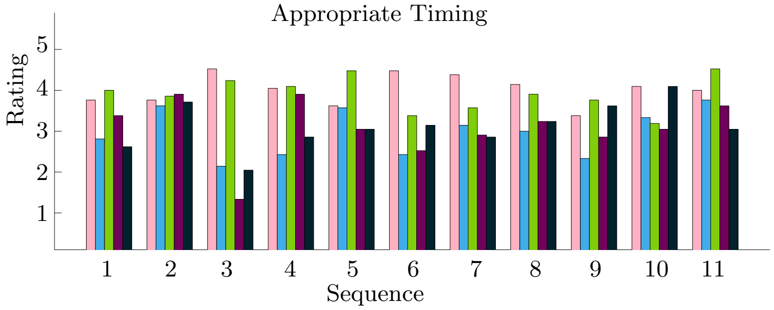

Fig. 5. Timing rating results from Experiment 1, by Sequence.

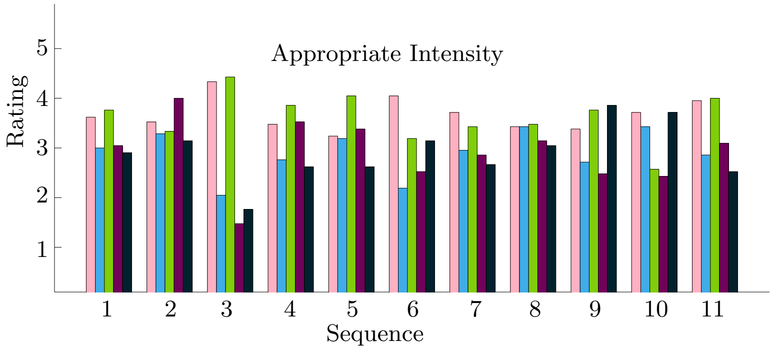

Fig. 6. Intensity rating results from Experiment 1, by Sequence.

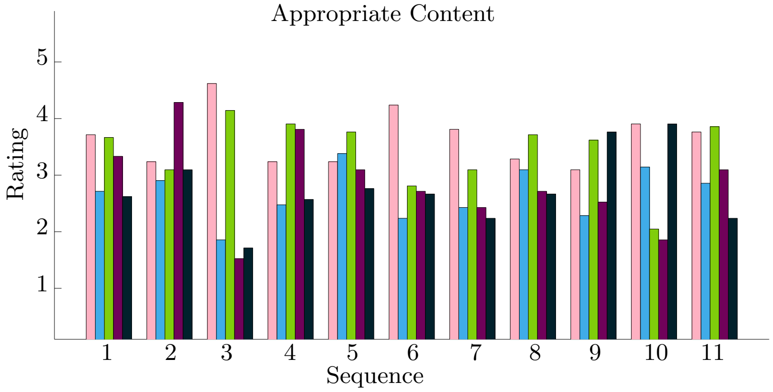

Fig. 7. Content rating results from Experiment 1, by Sequence.

3.4 Rating Performance

Figures 5–7 show the ratings for appropriateness of *timing*, *intensity*, and *content* (respectively) for all 55 backchannel clips (5 backchannels for 11 sequences). Overall, the pattern of results on the three rating scales is very similar – not only to each other, but also to the pattern of choices in the recognition task. For example, in Sequences 6 and 7, the original sequence was chosen very often in the recognition task and received high ratings on all three scales, while the remaining responses were chosen less often and received lower ratings. Likewise in Sequences 1 and 3, the 1st and 3rd alternatives were chosen very often and received proportionally higher ratings. Perhaps the biggest anomaly is Sequence 10, where alternatives 1 and 5 were rated equally high on all three scales, but alternative 5 was rarely chosen in the recognition task.

Sequences 2 and 8 prove to be the exceptions to the rule of rating similarity. In Sequence 2, alternative 4 was chosen most often while the other four alternatives were rarely chosen, which is reflected somewhat in the *intensity* ratings and very much so in the *content* ratings. All five alternatives, however, were rated equally appropriate in terms of *timing*. In Sequence 8, alternatives 1 and 3 were chosen rather often and received high *timing* and *content* ratings. In contrast, the five alternatives received similar *intensity* ratings.

The *humor* ratings diverge from the results of the other tasks quite a bit. Although there is a similarity between the *humor* ratings and the choice frequencies in the recognition task for Sequences 3, 7 and 8, the rest are either slightly anomalous or simply not diagnostic.

To further examine the relationship between performance on the tasks, we correlated all four rating dimensions with recognition performance to investigate any potential linear relationships in the data. Correlations were $r_{timing} = 0.73, r_{intensity} = 0.71, r_{content} = 0.77$, and $r_{humor} = 0.54$. Apart from the humor dimension, every rating dimension therefore carries some information about the recognition performance. We then conducted a linear regression with all four ratings as predictors on the recognition performance to see the contribution of each rating in a joint model. The resulting equation from the regression was: performance = 55 + timing × 8.65 + intensity × (-10.19) + content × 21.23 + humor ×4.45. The r^2 value for this model is $r^2 = 0.63$ indicating a good prediction performance. In this joint model, content carries the highest weight in predicting the recognition performance outcome. The absolute weight of both timing and intensity in the prediction are similar. Interestingly, intensity receives a negative weight, suggesting a reverse relationship. Finally, as could be seen already from the correlation data, humor has the lowest weight in the joint model.

4 Experiment 2: Sensitivity to Synchronization

Experiment 1 focused on investigating high-level, contextual sensitivity to backchannel responses. In Experiment 2 we focus on sensitivity to timing.

4.1 Methods

Stimuli In order to investigate sensitivity to timing, we chose to measure psychometric functions in a standard psychophysical experiment (using the method of constant stimuli). The baseline stimuli for this experiment consisted of the original 11 main-channel/backchannel sequences. For each of the 11 sequences, we created 6 more backchannel sequences – each the same length – by shifting the selection window backwards and forwards in time. We chose temporal offsets of -45 to +45 frames in 15 frame intervals (-1.5 to +1.5 seconds in 0.5 second intervals). The 77 sequences were repeated three times each in completely randomized order, yielding a total of 231 trials. This repetition is standard procedure in experimental design, see [15] for further details.

Procedure The experiment used the same hardware setup as in Experiment 1. A different set of 20 participants were recruited for this experiment. Once the participant was seated comfortably, the experiment began with instructions shown on-screen.

Each of the 231 main-channel/backchannel pairs was shown in random order to the participants. Participants were instructed to carefully watch the videos and to decide whether the backchannel response was too early or too late. The experiment lasted about 45 minutes. After the experiment, participants were paid for their participation, debriefed and any remaining questions were answered.

4.2 Results and Discussion

Psychometric functions were fitted to each participant's data using psignifit version 2.5.6, a software package which implements the maximum-likelihood method described in [16]. The fitted functions were used to derive two important psychophysical parameters: the point of subjective synchronicity (which is the time offset at which both main-channel and backchannel perceptually appear synchronized), and the just-noticeable-difference (which is the time difference between two sequence pairs that will be noticed as different). Note that the use of the fitted curve to derive the PSS and JND means that the JND can lie outside of the observed stimulus range. The data of four of the 20 participants proved to be anomalous (e.g, the thresholds diverged from the mean by more than one standard deviation).

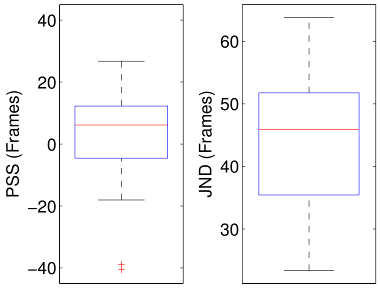

Fig. 8. Boxplots showing the distributions of (left) the point of subjective synchronicity and (right) the just-noticeable-difference.

The distribution of the point of subjective synchronicity (PSS) is shown in Figure 8). Its median is consistently around +6.11 frames . That is, participants felt that the backchannel properly matched the main-channel when the back channel lagged by 6 frames (about 200 milliseconds). Given the difficulty of the task, it seems that people are remarkably sensitive to the correct timing between a speaker and a listener.

One potential reason for the observed lag might lie in the high cognitive load imposed by the task. Usually when watching a conversation, we do not explicitly pay attention to the timing. By bringing this element of a conversation to conscious attention, participants may need additional cognitive resources (and thus additional time) to process the videos in a manner that is less automatic than usual. It is also possible that the delay is related to saccades as participants were required to redirect their gaze and attention from the speaker to the listener, when the listener begins to become more active (or is expected to become more active). It is well known that an intentional shift of gaze focus takes at least 200 ms [17].

The distribution of the just-noticeable-difference (JND) is shown in Figure 8. Its median is 45.9 frames, corresponding to 1.5s. This means that in order to reliably detect time offset differences between two sequences (in either direction), they would need to be shifted by 45 frames. The lowest JND among all participants was 0.7 seconds, the highest 2 second. Hence, whereas participants on average can detect the veridical time offset, their sensitivity to *changes* in synchronization is on the order of 1.5 seconds. Although this difference may seem large at first glance, one has to bear in mind that the JND did does not test the *detection* of backchannel responses, but rather measures how well *changes to the synchronization of two speakers* could be judged – something which imposes much more complex demands on conversation processing.

5 Conclusion

In this paper, we presented results of two experiments on contextual and temporal sensitivity to backchannel responses. Using a newly recorded database of natural conversations, we extracted several speaker-listener interactions containing non-verbal, visual backchannel responses (of the face, head, and shoulders) of a listener. In Experiment 1 we found that participants were well able to identify the correct backchannel response among a list of alternative responses – the success of this match, however, depended crucially on both speaker and listener. The additional dimensions that were analyzed correlated with recognition performance and we were able to predict recognition performance reasonably well using a joint linear model. A more detailed model using additional, important dimensions for conversational analysis, however, still needs to be investigated. In Experiment 2, we examined sensitivity to time offsets in the backchannel response. We found that participants' points of subjective synchronicity were on average almost veridical. Their sensitivity (as measured by the just-noticeable-difference) was around 1.5s, which is fairly good considering the complexity of the speaker-listener interaction. The two experiments here represent the start of our investigations into full spatio-temporal models of how facial expressions and facial gestures are used in conversational contexts. In future work, we will be constructing full active-appearance models of the speakers and listeners in the database. These will be used to create video sequences modified to freeze certain facial parts (*e.g.*, [7]), to warp the timing of the backchannel responses, *etc.* With these modified sequences, we can conduct more detailed experiments on the sensitivity to physical changes for facial expressions in conversational contexts.

References

1. Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., Schroeder, M.: Bridging the Gap between Social Animal and Unsocial Machine: A Survey of Social Signal Processing. IEEE Transactions on Affective Computing **3** (2012) 69–87
2. Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups: A review. Image and Vision Computing **27** (2009) 1775–1787
3. Ekman, P.: Universal and cultural differences in facial expressions of emotion. (1972) 207–283
4. Baron-Cohen, S., Wheelwright, S., Jolliffe, T.: Is there a "language of the eyes"? evidence from normal adults, and adults with autism or asperger syndrome. Visual Cognition **4** (1997) 311–331
5. Cunningham, D.W., Wallraven, C.: Dynamic information for the recognition of conversational expressions. Journal of Vision **9** (2009)
6. Pelachaud, C., Poggi, I.: Subtleties of facial expressions in embodied agents. The Journal of Visualization and Computer Animation **13** (2002) 301–312
7. Nusseck, M., Cunningham, D.W., Wallraven, C., Bülthoff, H.H.: The contribution of different facial regions to the recognition of conversational expressions. Journal of Vision **8** (2008)
8. Wehrle, T., Kaiser, S., Schmidt, S., Scherer, K.R.: Studying the dynamics of emotional expression using synthesized facial muscle movements. Journal of Personality and Social Psychology **78** (2000) 105–119
9. Heylen, D.: Challenges ahead: head movements and other social acts during conversations. In: Joint Symposium on Virtual Social Agents. (2005) 45–52
10. Bavelas, J.B., Coates, L., Johnson, T.: Listener responses as a collaborative process: The role of gaze. Journal of Communication **52** (2002) 566–580
11. Yngve, V.: On getting a word in edgewise. In: Papers from the Sixth Regional Meeting of the Chicago Linguistic Society. (1970) 567–578
12. Schröder, M., Heylen, D., Poggi, I.: Perception of non-verbal emotional listener feedback. In: Proceedings of Speech Prosody, Dresden, Germany (2006)
13. Poppe, R., Truong, K., Heylen, D.: Backchannels: Quantity, type and timing matters. In Vilhjálmsson, H., Kopp, S., Marsella, S., Thórisson, K., eds.: Intelligent Virtual Agents. Volume 6895 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2011) 228–239
14. Poppe, R., Truong, K.P., Reidsma, D., Heylen, D.: Backchannel strategies for artificial listeners. In: Proceedings of the 10th international conference on Intelligent virtual agents. IVA'10, Berlin, Heidelberg, Springer-Verlag (2010) 146–158
15. Cunningham, D.W., Wallraven, C.: Experimental Design: From User Studies to Psychophysics. A K Peters/CRC Press (2011)
16. Wichmann, F.A., Hill, N.J.: The psychometric function: I. fitting, sampling and goodness of fit. Perception and Psychophysics **63** (2001) 1293 – 1313
17. Trottier, L., Pratt, J.: Visual processing of targets can reduce saccadic latencies. Vision Research **45** (2005) 1349 – 1354