The MultiLis Corpus - Dealing with Individual Differences in Nonverbal Listening Behavior

Iwan de Kok and Dirk Heylen

Human Media Interaction
University of Twente
{koki, heylen}@ewi.utwente.nl

Abstract. Computational models that attempt to predict when a virtual human should backchannel are often based on the analysis of recordings of face-to-face conversations between humans. Building a model based on a corpus brings with it the problem that people differ in the way they behave. The data provides examples of responses of a single person in a particular context but in the same context another person might not have provided a response. Vice versa, the corpus will contain contexts in which the particular listener recorded did not produce a backchannel response, where another person would have responded. Listeners can differ in the amount, the timing and the type of backchannels they provide to the speaker, because of individual differences - related to personality, gender, or culture, for instance. To gain more insight in this variation we have collected data in which we record the behaviors of three listeners interacting with one speaker. All listeners think they are having a one-on-one conversation with the speaker, while the speaker actually only sees one of the listeners. The context, in this case the speaker's actions, is for all three listeners the same and they respond to it individually. This way we have created data on cases in which different persons show similar behaviors and cases in which they behave differently. With the recordings of this data collection study we can start building our model of backchannel behavior for virtual humans that takes into account similarities and differences between persons.

Key words: Multimodal corpus, listeners, task-oriented

1 Introduction

In the field of embodied conversational agents a lot of effort is put into making the interaction between humans and the agent as natural as possible. Models of interaction behavior can be based on insights from theoretical studies, rules of thumb or they can be based on machine analysis of multimodal corpora of human-human interactions. The problem with using multimodal corpora to analyze and model human behavior is that in the recorded face-to-face conversation only one example of an appropriate way to act in a certain context is recorded even though there may be other appropriate responses to that context than the

Fig. 1. Schematic overview of the setup for this corpus. Each interaction is between a speaker and three listeners. The *speaker* sees only one of the listeners (the *displayed listener*). The other two listeners, the *concealed listeners*, can see the speaker, but the speaker is unaware of their participation to the conversation. All listeners believe to be the only listener in the interaction.

one recorded. This certainly holds true for backchannel responses. In most situations, there is nothing wrong with missing a backchannel opportunity. When one is using a multimodal corpus to train a model for predicting the timing of backchannel continuers, as was done in [7], this optionality of behaviour provides a problem for evaluation. The issue of variability and optionality of behaviours show both in training and in evaluation. First, not all the contexts in which a backchannel is possible are used for analysis and modeling, only the contexts in which the listener recorded actually provided a backchannel response are included. Second, the issue also shows during evaluation of the model. Evaluation usually involves comparing the output of the backchannel model trained on the corpus to the actual responses recorded. There might be many cases in which the model predicts a possible backchannel which one might consider correct, but where the actual listener in the corpus did not produce one. These should not be counted as false positives but with most evaluations methods they will.

Besides in timing a backchannel listeners may differ in the type of backchannels they provide. Variation may be related to personality, gender, culture, or the time of day, concentration or attention levels, etcetera. To gain more insight in these differences and to help in training more accurate computational models, we have collected a new set of recordings that shows individual differences between multiple listeners that believe they are interacting face-to-face with the speaker. Each listener thinks he is the only one listening to the speaker and the speaker also thinks there is only one listener (see Figure 1). We will describe the MultiLis corpus in this paper.

Before we present the data collection proper, we will discuss related work which has attempted to deal with the same problem as stated above, in Section 2. The method of data collection we have performed in Section 3. In Section 4 the recordings of the corpus are presented and the annotations made are described in Section 5. The individual differences we captured in this corpus are shown in Section 6. Finally our plans for using this data are discussed in Section 7.

2 Related Work

We are not the first to identify this problem of optionality and to offer a solution. In their effort to create a predictive rule for backchannels based on prosodic cues, Ward and Tsukahara [10] define backchannel feedback as follows: “Backchannel feedback responds directly to the content of an utterance of the other, is *optional* [our emphasis], and does not require acknowledgement by the other.” When analyzing the performance of their predictive rule they conclude that 44% of the incorrect predictions were cases where a backchannel could naturally have appeared according to the judgement of one of the authors, whereas in the corpus there was silence or, more rarely, the start of a turn. Listeners might have given a nonverbal backchannel (e.g., a head nod), which is not recorded, but the listeners may also have chosen not to produce a backchannel at all at that opportunity.

Cathcart et al. [2] identified the same problem as well. In their shallow model of backchannel continuers based on pause duration and n -gram part -of-speech tags they remark that human listener differ markedly in their own backchanneling behavior and pass up opportunities to provide a backchannel. Their attempt at dealing with this is to test their model on high backchannel rate data which is based on the assumption that the more backchannels an individual produces, the fewer opportunities they are likely to have passed up.

Both the papers attempt to solve the problem after the data is collected and the model has been produced. This gives them a better insight into the actual qualitative performance of the model by taking the variation into account, but it does not solve the issue at its root, namely the way in which ground truth is established for the contexts in which backchannels are appropriate.

Noguchi and Den [8] look at an alternative way to collect data. In their machine learning approach to modelling back-channel behavior based on prosodic features they need a collection of positive and negative examples of appropriate contexts for backchannels. Because providing backchannels is almost always optional, they also argue that it is not appropriate to consider only those contexts where backchannels are found in the corpus as positive examples and contexts where no backchannels are found as negative examples. As a solution to this they collected backchannels responses from participants in a study, in which the participants looked at a video of a speaker and were asked to hit the space bar on a keyboard at times where they thought a backchannel response was appropriate. The stimuli consisted of several pause-bounded phrases which constitute a single conversational move. Each stimulus was shown to 9 participants. By counting the number of participants that responded positively to the stimuli they classi-

fied each one as either an appropriate context for a backchannel, an inappropriate context for a backchannel or as indecisive.

The same approach was also used by Huang et al. [4] where it was dubbed Parasocial Consensus Sampling (PCS). The authors used the annotation thus obtained directly to generate backchannel responses of a virtual human in a simulated conversation with the original speaker. In a subsequent evaluation study, participants rated these conversations and compared them to conversations between the human speaker and a virtual agent that was copying the original listener behavior. Participants thought the agent that used the consensus-based algorithm was more believable and showed higher rapport with the speaker than the agent that copied the original listener.

The Parasocial Consensus Sampling method is particularly useful for analyzing the timing of backchannels, however, there are many more dimensions of human listening behavior that should be accounted for. People also vary in the selection of the type of backchannel they choose (e.g., only a head nod, a head nod accompanied by 'yeah', a head nod accompanied by 'mm-hmm', etc.), or their gaze behavior and body posture. Using the PCS method to annotate these features as well is very time consuming, if not impossible for some behaviors (e.g., body posture). Therefore we propose another way of collecting a corpus for the analysis of (non-)verbal listening behavior which accounts for variation between listeners.

3 Data Collection

The goal of this data collection is to record multiple listeners in parallel in interaction with the same speaker. The corpus consists of interactions between one speaker and three listeners in which each of the listeners sees the speaker on a monitor and the speaker sees only one of the listeners. The listeners are made to believe they have a one-on-one conversation with the speaker and also the speaker is unaware of the special setup. How this illusion is created is discussed in Section 3.1. The procedure during recording is discussed in Section 3.2, the tasks are discussed in Section 3.3 and the measures we used can be found in Section 3.4.

3.1 Setup

Each of the participants sat in a separate cubicle. The digital camcorders which recorded the interaction, were placed behind a one-way mirror onto which the interlocutor was projected (see Figure 2). This ensured that the participants got the illusion of eye contact with their interlocutor. In Figure 3 one can see that the listeners appear to be looking into the camera which was behind the mirror. This video was also what the participants saw during the interaction. All participants wore a headphone through which they could hear their interlocutor. The microphone was placed at the bottom of the autocue set up and was connected to the camcorder for recording.

Fig. 2. Picture of the cubicle in which each participant was seated. It illustrates the interrogation mirror and the placement of the camera behind it which ensures eye contact.

During the interaction speakers were shown one of the listeners (the *displayed listener*) and could not see the other two listeners (the *concealed listeners*). All three listeners saw the recording of the same speaker and all three believed to be the only one involved in a one-to-one interaction with that speaker. Distribution of the different audio and video signals was done with a Magenta Mondo Matrix III, which is a UTP switch board for HD-video, stereo audio and serial signals. Participants remained in the same cubicle during the whole experiment. The Magenta Mondo Matrix III enabled us to switch between distributions remotely.

3.2 Procedure

In total eight sessions were recorded. For each session there were four participants invited (in total there were 29 male and 3 female participants, with a mean age of 25). At each session, four interactions were recorded. The participants were told that in each interaction they would have a one-on-one conversation with one other participant and that they would either be a speaker or a listener. However, during each interaction only one participant was assigned the role of speaker and the other three were assigned the role of listeners. Within a session, every participant was a speaker in one interaction, was once a displayed listener and appeared twice as concealed listener.

In order to be able to create this illusion of one-on-one conversations we needed to limit the interactivity of the conversation because as soon as the displayed listener would ask a question or start speaking, the concealed listeners would notice this in the behavior of the speaker and the illusion would be broken. Therefore the listeners were instructed not to ask questions or take over the role

of speaker in any other way. However we did encourage them to provide short feedback to the speaker.

3.3 Tasks

The participants were given tasks. There were two different type of tasks the participants needed to perform during the interaction, either retelling a video clip or giving the instruction for a cooking recipe. For the retelling of the video speakers were instructed to watch the video carefully and to remember as many details as possible, since the listener would be asked questions about the video after the summary of the speaker. To give the speakers an idea of the questions which were going to be asked, they received a subset of 8 open questions before watching the video. After watching the video they had to give the questions back so that they would not have something to distract them. After the retelling both the speaker and the listeners filled out a questionnaire with 16 multiple choice questions about the video. Each question had four alternative answers plus the option “I do not know” and for the listener the extra option “The speaker did not tell this”. As stimuli the 1950 Warner Bros. Tweety and Sylvester cartoon “Canary Row” and the 1998 animated short “More” by Mark Osborne were used¹.

For the second task the speaker was given 10 minutes to study a cooking recipe. Both the listener and the speaker needed to reproduce the recipe as completely as possible in the questionnaire afterwards. As stimuli a tea smoked salmon recipe and a mushroom risotto recipe were used.

We chose to use two different tasks to be able to see the influence of the task on listening behavior. The retelling of the video is more entertaining and narrative in nature, while the recipe task is more procedural.

3.4 Measures

Before the experiment we asked participants to fill out their age and gender and we had them fill out personality and mood questionnaires. For personality we used the validated Dutch translation of the 44 item version of the Big Five Inventory [6]. For mood we used seven out of eleven subscales from Positive and Negative Affect Schedule - Expanded Form (PANAS-X, 41 items) [11] and the two general positive and negative affect scales. Furthermore we used the Profile of Mood States for Adults (POMS-A, 24 items) [9]. For both PANAS-X and POMS-A we used unvalidated Dutch translations made by the authors. Subjects were instructed to assess their mood of “today”.

After each interaction speakers filled out the Inventory of Conversational Satisfaction (ICS, 16 items) [12], questions about their task performance (5 items) and questions about their goals during the interaction (3 items). The listeners

¹ Canary Row (1950): <http://www.imdb.com/title/tt0042304/>; More (1998): <http://www.imdb.com/title/tt0188913/>.

Fig. 3. Screenshot of a combined video of the four participants in an interaction.

filled out an adapted version of the rapport measure [3] with additional questions from the ICS (10 items in total, e.g. "There was a connection between the speaker and me."). Some questions of the 16 items ICS relate to talking, which the listener does not do in our experiment, so they were left out. Furthermore the listeners answered six questions about the task performance of the speaker, such as "The speaker was entertaining" or "The speaker was interested in what he told". All questions were 5-point Likert Scale.

After the complete session, with all four interactions were finished subjects were debriefed and were asked which interaction they preferred; whether they had believed the illusion of always having one-on-one interaction, and if not, at which moment they had noticed this; in which interaction they thought the speaker could see them; about the delay of the mediated communication, audio and video quality (3 items).

4 MultiLis Corpus

In total 32 conversations were recorded (8 for each task), totalling in 131 minutes of data (mean length of 4:06 minutes). All the conversations were in Dutch.

Audio and video for each participant was recorded in synchrony by the digital camcorders. Synchronisation of the four different sources was done by identifying the time of a loud noise which was made during recording and could be heard on all audio signals.

Videos are available in high quality (1024x576, 25fps, FFDS compression) and low quality (640x360, 25fps, XviD compression). Audio files are available in high quality (48kHz sampling rate) and low quality (16kHz sampling rate). Furthermore a combined video (1280x720, 25fps, XviD compression) of all four participants in a conversation is available (for a screenshot, see Figure 3).

5 Annotations

Speakers were annotated on eye gaze and mouth movements other than speech. Listeners were annotated on head, eyebrow and mouth movements and any speech they produced was transcribed as well. For this annotation we used the ELAN annotation tool [1].

For the listeners the annotations were made in a three step process. First the interesting regions with listener responses were identified. This was done by looking at the video of the listener with sound of the speaker and marking moments in which a response of the listener to the speaker is noticed. In the second step these regions were annotated more precisely on head, brows and mouth movements. Speech of the listener was also transcribed by hand. In the third and final step the onset of the response was determined.

In the following subsections the annotation scheme for each modality will be explained in more detail. In each annotation scheme left and right are defined from the perspective of the annotator.

EYE GAZE Annotation of the speakers' gaze provides information about whether they were looking into the camera (and therefore looking at the listener) or not and whether there was blinking. For each of these two features a binary tier was created. Annotations were done by two annotators who each annotated half of the sessions. One session was annotated by both. Agreement (calculated by overlap / duration) for gaze was 0.88 and for blink 0.66.

HEAD For listeners the shape of the head movements. An annotation scheme of 12 categories was developed. The 12 categories and the amount of annotations in each category are given below. Several movements had a lingering variant. Lingering head movements are movements where one there is one clear stroke identifiable followed with a few more strokes that clearly decrease in intensity. If during this lingering phase the intensity or frequency of the movement increases again a new annotation is started.

- **Nod** (681 & 766 lingering) - The main stroke of the vertical head movement is downwards.
- **Backnod** (428 & 290 lingering) - The main stroke of the vertical head movement is upwards.
- **Double nod** (154 & 4 lingering) - Two repeated head nods of the same intensity.
- **Shake** (17) - Repeated horizontal head movement.
- **Upstroke** (156) - Single vertical movement upwards. This can either occur independently or just before a nod.
- **Downstroke** (43) - Single vertical movement downwards. This can either occur independently or just before a backnod.
- **Tilt** (24 left & 15 right) - Rotation of the head, leaning to the left or right.
- **Turn** (8 left & 11 right) - Turning of the head into left or right direction.
- **Waggle** (7) - Repeated nods accompanied by multiple head tilts.

- **Sidenod** (9 & 2 lingering) - Nod accompanied by a turn into one direction (6 left & 5 right).
- **Backswipe** (18 & 2 lingering) - Backnod which is not only performed with the neck, but also the body moves backwards.
- **Sideswipe** (3 left & 5 right) - Sidenod which is not only performed with the neck, but also the body moves into that direction.

Keep in mind that head movements are annotated only in areas where a listener response was identified in the first step of the annotation process. Especially turns and tilts occur more often than reflected in these numbers, but the others are not categorized as listener responses.

EYEBROWS For the listeners eyebrow raises and frowns were annotated. It was indicated whether the movement concerned one or both eyebrows. When one eyebrow was raised or frowned, it is indicated which eyebrow (left or right) made the movement. In total this layer contains 200 annotations, 131 raises and 69 frowns. These numbers include the annotations in which only one eyebrow was raised or frowning occurred with one eyebrow.

MOUTH The movements of the mouth are annotated with the following labels (457 in total): smile (396), lowered mouth corners (31), pressed lips (22) and six other small categories (8). Especially with smiles the end time is hard to determine. If the person is smiling, but increases the intensity of the smile, a new annotation is created.

SPEECH For the speakers we collected the results of the automatic speech recognition software SHoUT [5]. For listeners the speech was transcribed. In total 186 utterances were transcribed. The most common utterances were “uh-huh” (76), “okay” (42) and “ja” (29).

RESPONSES This annotation layer was created in the third step of the annotation process of the listener. What we refer to as a listener response can be any combination of these various behaviors, for instance, a head nod accompanied by a smile, raised eyebrows accompanied by a smile or the vocalization of uh-huh, occurring at about the same time. For each of these responses we have marked the so-called onset (start time). The onset of a listener response is either the stroke of a head movement, the start of a vocalization, the start of eyebrow movement or the start of a mouth movement. When different behaviors combine into one listener response, either the head movement or vocalization was chosen as onset (whichever came first). If there was no head movement or vocalization present, either the eyebrow or mouth movement was chosen as onset (whichever came first). In total 2796 responses are in the corpus.

Given these annotations, we can start to look at the distribution of the various responses of one listener, both in relation to the speaker’s behaviours and in relation to the responses of the other listeners.

Noticed illusion?		Seen by speaker?		
		Correct	No Idea	Wrong
Yes	14	6	3	5
No	17	5	1	11
All	31	11	4	16

Table 1. The table shows the reported answers to questions in the post-questionnaire regarding the manipulation. 14 out of 31 participants reported that they noticed that they were not always seen. In total 11 out of 31 participants could correctly identify the interaction in which they were seen by the speaker.

Fig. 4. Histogram showing the distribution of speakers in percentage of speaking time.

6 Individual Differences in the MultiLis Corpus

The reason for collecting this corpus was to identify and be able to analyze the differences and similarities between individuals in their listening behavior. In this section we show that we were successful in capturing these differences and similarities in the corpus. We start with a manipulation check to see whether our illusion of one-on-one conversation instead of parallel conversation worked. After this we show results illustrating the differences between the speaker and the listeners captured in the corpus.

Table 1 shows the results of the manipulation check. During the post-questionnaire 14 out of 31 participants reported that they noticed that they were not always seen (1 participant failed to fill out the post-questionnaire). Out of those 14 only 6 could correctly identify in which conversation they were seen by the speaker. In total 11 out of 31 participants correctly identified this conversation. This is hardly above chance ($= 33\%$). So, even when they noticed something was off, they could still not identify the correct conversation. If we look at the amount of responses the displayed listeners (1007) and the concealed listeners (895) gave we see that concealed listeners give on average 12% less responses, but according to a paired-sample t -test between the amount of responses per minute per session of displayed listeners and concealed listeners this difference is not significant ($p = 0.33$).

Fig. 5. Histogram showing the distribution of speakers in percentage of gaze directed at the listener.

Fig. 6. Histogram showing the distribution of speakers in gaze shifts per minute.

An analysis of the speakers' annotations tells us that speakers differ in the amount of time they spoke (Figure 4), the amount of time they directed their gaze towards the listener (Figure 5) and the amount of gaze shifts (Figure 6). Speaking time varied between speaking 46.8 % to 83.1% of the time (mean 67.9%). The variation in gaze directed at the listeners is bigger with percentages ranging from 16.3% to 97.0% (mean 67.3%). Also the amount of gaze shifts differs a lot between speakers, from 1.6 to 24.7 gaze shift per minute (mean 13.3 per minute).

Figure 7 shows the amount of responses the listeners gave. For this histogram the average amount of responses of the three interactions of each listener is used. The amount of responses varied between 2.5 and 15.9 responses per minute (mean is 7.5 responses per minute).

Each listener was involved in three interactions. When the amount of responses the listener provides was solely caused by the choices and preferences of the listener, independently of the speaker, the listener would provide on average the same amount of responses in each of these three interactions. Figure 8 shows that this is not the case. We have grouped the interactions with the least amount of responses of each listener, the interactions with the most responses and the ones in between and calculated the mean responses per minute for each

Fig. 7. Histogram showing the distribution of listeners in responses per minute

group. The right graph in Figure 8 shows that for the group the amount of the interactions with the least responses is 3.9 responses per minute, for the middle group 6.9 responses and the group with the most response, 10.5 responses per minute. So, there is a significant variation within the behavior of the listener, caused by the speaker ($p < 0.001$ on a paired-sample t -test between all groups).

The speaker is not the sole factor here as well. Figure 8 also shows that there is a variation between the three listeners within a session. Calculating the mean responses per minute for each group of the listeners of each session with the least responses per minute ($= 4.8$), most responses ($= 9.4$) and the group in between ($= 7.0$), we get the left graph in Figure 8. The variation is not as big, but still significant ($p < 0.001$ on a paired-sample t -test between all groups). This suggests that the combination of speaker and listener determines the amount of responses the listener gives, where the speaker has a little more influence on this.

To get an indication about the times that multiple listeners produced a response at about the same time, we clustered together the responses of the three listeners. Leaving out smiles and brow movements we identified 1735 clusters. In 128 cases, all three listeners produced a listener response and in 456 cases there were two listeners responding at about the same time. In 1142 of the cases only one of the listeners produced a response.

7 Conclusions and Future Work

With this data set we have created a rich source for researching various aspects of listener behavior. The corpus especially provides more insight into individual differences and similarities in nonverbal listening behavior than previous data sets, because of its unique setup. The setup deals with the limitation of previous corpora of only recording one example of an appropriate way to act in a certain context, by recording three listeners in parallel. With this data intend to improve the learning and evaluation of prediction or generation models for nonverbal listening behavior.

Fig. 8. The left figure shows the average amount of responses per minute grouping together the three listener *within a session* into a group of listeners with the least responses, the listeners with the most responses and the listeners in between. This illustrates that within the three listeners the amount of responses varies quite a lot. The right figure shows the average amount of responses per minute when looking at the three sessions of *each individual listener* and when grouping them into the sessions with the least responses, the listeners with the most responses and the listener in between. This illustrates that the response behavior of each listener is highly influenced by the actions of the speaker.

Acknowledgments

We would like to thank Khiet Truong, Ronald Poppe and Alfred de Vries for giving a helping hand during the collection of the corpus, Hendri Hondorp for his help in synchronizing all recordings and Arie Timmermans for his help annotating the data. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 211486 (SEMAINE).

References

1. Brugman, H., Russel, A.: Annotating multimedia/multi-modal resources with ELAN. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation. pp. 2065–2068. Citeseer (2004)
2. Cathcart, N., Carletta, J., Klein, E.: A shallow model of backchannel continuers in spoken dialogue. European ACL pp. 51–58 (2003)
3. Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R.: Creating rapport with virtual agents. In: Proceedings of Intelligent Virtual Agents. pp. 125–138. Springer-Verlag New York Inc, Paris, France (2007)
4. Huang, L., Morency, L.P., Gratch, J.: Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behavior. In: Proceedings of Autonomous Agents and Multi-Agent Systems. Toronto, Canada (2010)
5. Huijbregts, M.: Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled. Phd thesis, University of Twente (2008)
6. John, O.P., Naumann, L.P., Soto, C.J.: Paradigm shift to the integrative Big-Five trait taxonomy: History, measurement, and conceptual issues, chap. 4, pp. 114–158. Guilford Press, New York, New York, USA, 3 edn. (2008)

7. Morency, L.P., de Kok, I., Gratch, J.: A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems* 20(1), 70–84 (May 2010)
8. Noguchi, H., Den, Y.: Prosody-based detection of the context of backchannel responses. In: *Fifth International Conference on Spoken Language Processing* (1998)
9. Terry, P.C., Lane, A.M., Fogarty, G.J.: Construct validity of the Profile of Mood States-Adolescents for use with adults. *Psychology of Sport and Exercise* 4(2), 125–139 (2003)
10. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics* 32(8), 1177–1207 (July 2000)
11. Watson, D., Clark, L.A.: *The PANAS-X* (1994)
12. White, S.: Backchannels across cultures: A study of Americans and Japanese. *Language in Society* 18(1), 59–76 (1989)