Twente Debate Corpus – A Multimodal Corpus for Head Movement Analysis

Bayu Rahayudi, Ronald Poppe, Dirk Heylen
University of Twente
PO Box 217, 7500AE, Enschede, Netherlands
E-mail: b.rahayudi@utwente.nl, r.w.poppe@utwente.nl, d.k.j.heylen@utwente.nl

Abstract

This paper introduces a multimodal discussion corpus for the study into head movement and turn-taking patterns in debates. Given that participants either acted alone or in a pair, cooperation and competition and their nonverbal correlates can be analyzed. In addition to the video and audio of the recordings, the corpus contains automatically estimated head movements, and manual annotations of who is speaking and who is looking where. The corpus consists of over 2 hours of debates, in 6 groups with 18 participants in total. We describe the recording setup and present initial analyses of the recorded data. We found that the person who acted as *single* debater speaks more and also receives more attention compared to the other debaters, also when corrected for the time speaking. We also found that a *single* debater was more likely to speak after a *team* debater. Future work will be aimed at further analysis of the relation between speaking and looking patterns, the outcome of the debate and perceived dominance of the debaters.

Keywords: Debates, Multimodal corpus, Head movement

1. Introduction

The automatic analysis of behavior in face-to-face interactions has resulted in a better modeling and understanding of the communication process and it presents opportunities for technological support (Vinciarelli et al., 2012). Social signal processing is the research field that focuses on the social aspects of the automatic analysis. So far, much effort has been devoted to the analysis of facial expressions, body movement and vocal behavior but head movements have received much less attention (Heylen, 2006). Their importance in the turn-taking process and close relation with speech makes them important for the study of human face-to-face interactions.

Especially in multi-party face-to-face settings, head movements are intrinsically part of the turn-taking process, and have been found to be good predictors for the start of a new turn, listener comprehension (Battersby & Healey, 2010; Hadar, Steiner, & Clifford Rose, 1985) and participant role in the interaction (Salamin & Vinciarelli, 2012).

An interesting application is the analysis of debates, in which participants not only try to convince each other and respond to each other with words, but also nonverbally. While the automatic analysis of arguments in debates has received some attention (Bohus & Horvitz, 2011a; Pesarin et al., 2012; Salamin & Vinciarelli, 2012; Verbree, Rienks, & Heylen, 2006), the research on the nonverbal aspects of debates is limited. Bousmalis, Mehu, & Pantic (2013) address detecting agreement and disagreement from nonverbal cues. While they also focus on the temporal aspect of the detection, (dis)agreement is just one aspect of the debating process. Persuasion, cooperation and dominance all have an influence on the outcome of a debate, see also Curhan & Pentland (2007). Of particular interest is how these aspects play a role in debates in which multiple people participate that have either shared or conflicting views.

To the best of our knowledge, the analysis of head movements in debates with the aim of understanding and predicting debate outcome and the role of each participant in it, has not received any attention. We attribute this, at least partly, to the lack of suitable corpora.

In this paper, we present the Twente Debate Corpus (TDC), a novel multimodal corpus of three-person debates. In each session, one participant alone debates against a team of two participants. The *single* debater has an opposed opinion about the debated statement. This setting allows us to look at differences in communication patterns between participants with shared views and goals, and those with conflicting views. The corpus is suitable for the study into head movement and speech behavior for competing and cooperating debaters. Moreover, it can be used to study how cooperation affects the outcome of the debate.

We will first discuss related work on head movement analysis in multi-party settings, and existing multimodal databases that target this setting. Our corpus is introduced in Section 3. In Section 4, we will present the results of initial analyses. We conclude with a discussion of ways to use the corpus.

2. Related Work

Head movements have been found to correlate with gaze, which serves, amongst others, as a signal of attention (Heylen, 2006; Vertegaal, et al., 2001). Several studies have addressed the interpretation of head movements in the context of an interaction, in particular aimed at understanding and modeling their role in the turn-taking process (Bohus & Horvitz, 2011b; Hadar et al., 1985). Nods and shakes have received some attention, mainly due to their discrete nature and function as listener responses (de Kok, et al., 2010). Often, speech is also considered when studying head movements. Listener responses are strongly tied to the speech of the speaker (Ishi, Ishiguro, & Hagita, 2013). Moreover, patterns in head movement in both the speaker and listener have been found in the turn-taking process. Duncan (1972) observed that speakers turn away their head at the start of an utterance, while they turn to their interaction partner as a way to hand over the turn. Moreover, given its importance as a signal of attention, head movements of the speaker have been found a strong cue for the recognition of the addressee of the utterance (Jovanovic & op den Akker, 2004).

There are several multimodal corpora for the study of head movements in face-to-face settings. The AMI Meeting Corpus contains multimodal recordings of meetings in which four participants have to design a piece of equipment (Carletta, 2007). Manual and (semi-) automatic annotations on various levels are provided. Each participant acts out a specific role. The corpus is rich in the type of conversation but the participants' goals are shared and the acting reduces the naturalness of the interaction. Moreover, head movements were labeled afterwards which introduces issues with reliability. Recently, the Cardiff Conversation Database was introduced to study nonverbal communication, specifically backchannel behavior, in two-person interactions (Aubrey et al., 2013). While the interactions were natural and much attention was paid to the analysis of the face and head, the corpus contains neither multi-party interactions nor discussions. Similarly, the MHi-Mimicry database contains head movements but only in dialogs (Sun et al., 2011). The recording of head movements was explicitly addressed in their setting. The interactions concerned debatable issues but the views of the participants often quickly converged as they would be defending arbitrary views.

One multimodal corpus that explicitly concerns debates in a multi-party setting is the Canal9 Political debate database (Vinciarelli, et al., 2009). The recordings are taken from a television channel and contain heated debates between politicians. As the recordings are edited for broadcasting, there are many changes in camera view. Many times, only one of the debaters is shown, which means that no head movement data is available for the others.

The purpose of the current corpus is to target a multi-party setting in which both competition and cooperation in discussions can be studied, in particular in relation to the head movements of the participants.

3. Data Collection

In this section, we describe the scenario, recording, post-processing and annotation of the Twente Debate Corpus (TDC). The database (including the annotations, and head movement data) will be made publicly available through the web as a shared resource for the community, to develop features, classifiers and conversational models.

3.1 Scenario and Procedure

In the corpus, three participants debate statements of which one participant has an opposed opinion compared to the other two. The participants should convince the others of their view. The two participants are to act as a team. The debates are face-to-face discussions in which participants have equal views of the others.

Upon entering the recording room, participants were asked to sign a consent form, and filled in a 10-item Big-Five personality questionnaire (Gosling, Rentfrow, & Swann, 2003). They also filled in their opinion (agree, neutral or disagree) on a set of 35 debatable statements. Then, they engaged in a short interaction in which they got acquainted and familiarized themselves with the recording environment.

Based on the questionnaire with debatable items, three statements were selected. In each of these, one of the subjects did not agree with the other two. Based on these opinions, we then had three sessions in which a different subject was the *single* debater, and the others were a *team*. Sessions ended when the debate became repetitive, when consensus was reached or after 8 minutes, whichever came first. After each session, the participants were asked to fill in a questionnaire about the discussion. They were asked how convincing they found themselves and the others, and to indicate the level of cooperation within the team. The total duration of a session was approximately 40 minutes.

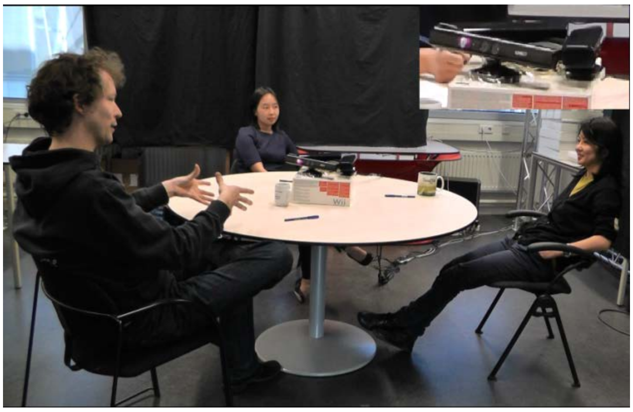

Figure 1: physical setup of the corpus recordings

3.2 Recording

The three participants were seated around a round table, in such a way that each of them faced the others at an equal angle. In the middle of the table there were three Microsoft Kinects (inset in Figure 1) that recorded the participants' video and head movement. In addition, we recorded speech with a microphone that was placed between the Kinects. Two computers were used to record the Kinect and microphone signals. There was also a camcorder in the corner of the room that recorded the entire setting (see Figure 1) and was used to facilitate the synchronization of the different recordings.

Every participant was recorded with a front view, at a 640x480 pixel resolution, 20 frames per second. Recordings were saved as MPEG-4 files. The Microsoft Kinect SDK was used to obtain head movements (pan, tilt, roll and the 3D head position relative to the Kinect) of each participant. An example of the head rotation of one of the participants can be seen in Figure 2.

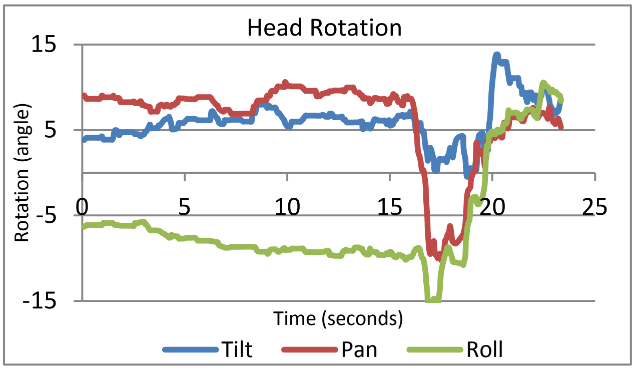

Figure 2. Example head rotation for one participant

The audio was recorded using a microphone that was wirelessly connected to a computer, and recorded using Audacity software and saved as 44.1 kHz 16-bit WAV file.

3.3 Post-processing

Given that the recordings had been made on different computers, they had to be aligned and synchronized in time. We included a clap in the recordings at the start of each session to facilitate this process. The Kinect video and head movement recordings were already synchronized as they were recorded with the same software program. We aligned these recordings and the audio to the overview camera video. We then cut the recordings into segments, each corresponding to one session. The timestamps of the head movement data were recalculated from the start of the segment.

The Kinect head movement software was not able to estimate the rotation and position of the head in every frame. This was typically the case when the hand moved in front of the head. Lack of head movement estimates results in gaps in the recordings. We decided to linearly interpolate gaps covering an interval shorter than two seconds. Larger gaps were left unchanged. From the 6 groups, we could only process 5 groups of data, because there was no head movement data recorded for one participant in one group. In total, 12 % of the corpus data did not have head movement data. For the analyses in this paper, we do not use the Kinect data so all sessions are included.

3.4 Annotation

ELAN (Brugman & Russel, 2004) was used to annotate the corpus on two distinct layers: “speaking” and “visual focus of attention” (VFOA). The annotations were made by the first author.

3.4.1 Speaking

It was annotated when a participant started or stopped speaking. The segment between a start and stop was labeled *speaking*, the segments between a stop and start as *not speaking*. Utterances with pauses less than one second were not considered as a different segment, but treated as the same segment. Given that we look at turns, and not at backchannels, such a treatment makes sense. As the annotations were made for each participant individually, pauses and overlap can be analyzed by considering the annotations of the three participants simultaneously.

3.4.2 Visual Focus of Attention

VFOA refers to the annotation of who is looking at whom at each moment in time. The annotation had four potential labels: *A*, *B*, *C* and *other*. The former three correspond to the three participants in the session. A person could not be looking at himself, which means that for each subject, there could be three distinct labels including *other*. The *other* class was used to annotate whenever a participant was not looking to other participant, but looking at the table, ceiling, camera or any other non-participant target. Given that we are interested in turn-taking behavior, we did not deem it important to distinguish between different other targets. Participants were not instructed regarding their behavior. From the automatic head position and rotation recordings, we could determine the VFOA labels as well, but only for those moments where no data is missing.

4. Corpus Statistics

The corpus contains 6 groups of debates, each consisting of 3 sessions. Each session contains 3 recordings for each participant. The total length of the 54 recording is 372.39 minutes, or 6.54 minutes on average. There were 18 participants (13 male and 5 female). Most of the participants were staff or student at the University of Twente. Their ages were between 22 and 60 years (30 years on average). All conversations were in English, although none of the speakers was a native English.

4.1 Speaking characteristics

From the annotation data, we know who is speaking when. We calculated the percentage of speaking when the participant acted as a *single* or as part of a *team*. We expect more speech from a *single* debater compared to a debater cooperating in a *team*, as they will require time to express their opinion and react to that of the *team* debaters. Moreover, we expect that the two *team* debaters together will speak more than the *single* debater, as they will add to each other’s statements. Indeed, the *single* debater talks on average 45.35% of the time, compared 27.52% for a *team* debater. This is approximately 1.65 times more, as expected, but less than the two *team* debaters together.

There were 627 *speaking* and 643 *not speaking* segments, with an average length of 13.85 and 30.03 seconds, respectively. When the participant acted as *single* debater she/he spoke for 15.32 seconds on average, compared to 12.95 seconds for a *team* debater. For *not speaking* segments, these durations are 19.13 and 35.47 seconds, respectively. These numbers indicate that *single* debaters not only speak more, but also speak more frequently.

We therefore expect that there is a pattern in the debates in which the single debater more often responds to one of the team debaters, than a team debater responds to her/his partner. To investigate whether this is the case, we analyzed the speaking turn sequences. There are three options for who speaks after whom: *single/team* (*single* debater followed by *team* debater), *team/single* and *team/team* (*team* debater followed by partner). The occurrence of these options is summarized in Figure 3.

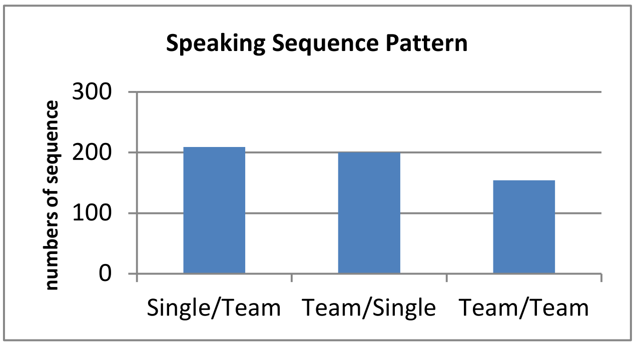

Figure 3. Comparison between *single/team, team/single*. and *team/team* turn patterns

In 209 cases, the *single* debater is followed by one of the *team* debaters, and in 200 the opposite is the case. Both these patterns appear more often than a *team* debater speaking after her/his partner, which happens 154 times. Apparently, the debaters in favor and against a certain statement speak more often in an alternating manner. These numbers are further supported by the fact that a *single* debater talks more and more frequently.

4.2 Looking characteristics

From the annotation process, we also obtain information who is looking at whom, and when this happens. We can calculate the percentage of looking at, and being looked at by other participants when a participant is a *single* or a *team* debater. We expect that a *team* debater will look more at the *single* debater than at her/his partner. This will be partly the case because the *single* debater speaks more but also more often is the addressee. These numbers are summarized in Table 1.

		Looked at by	
		Single	Team
Looks at	Single	-	56.28%
	Team	39.02%	27.64%

Table 1. Percentage of VFOA for *single* and *team*

While both a *single* and *team* debater spend approximately 80% of the time looking at other participants, there is a difference in the amount of visual attention that they get. A *single* and *team* debater was looked at by the other two on average 56.28% and 33.33% of the time, respectively. Our expectation is confirmed as the *single* debater indeed receives more attention – is looked at approximately 1.9 times more by the others – than a *team* debater. To analyze whether the difference in speaking time is of influence, we combine the speaking and looking annotations.

4.3 Speaking and looking combined

We also analyzed how long a person was looked at by the others on average when he acted as *single* and *team* when he was *speaking* and *not speaking*.

We can see from Table 2, that when a *single* debater was the speaker, he would be looked at by the other participants (*team*), on average 73.3% of an utterance. But when one of *team* members acted as the speaker, then he would be looked at by the other *team* member on average 50.3% of an utterance period, and be looked at by a *single* on 70.48%.

		Looked at by	
		Single	Team
Looks at	Single	-	73.3%
	Team	68.7%	50.3%

Table 2. Percentage of 'looking-at' for *single* and *team*, when speaking

		Looked at by	
		Single	Team
Looks at	Single	-	41.6%
	Team	21%	16%

Table 3. Percentage of time of 'looking-at' for *single* and *team*, when not speaking

From Table 3, we see that when the *single* debater was not speaking, he would be looked at by the other participants (*team*) on average 41.6% of an utterance. A non-speaking *team* member would be looked at by the other member *team* member on average 16% of an utterance, and looked at by *single* on 21%.

From Table 2 and Table 3, we can summarize that when the participant acted as *single*, he received more attention, both when he was speaking and when he was not. This can be explained by the fact that she/he was more often the addressee of the utterance. Moreover, it is to be expected that both *team* debaters would monitor the responses of the *single* debater, and anticipated that she/he would speak next.

5. Conclusion and Future Work

This paper introduced the Twente Debate Corpus, a three-person face-to-face debate corpus, which contains videos of each participant, audio of the debate, automatically analyzed head movements and annotations of who is speaking and visual focus of attention. A participant either acted individually, or with another participant as a team. The *single* and *team* debaters had conflicting views on debatable statements.

A total of over 2 hours of debate was recorded in 6 different sessions, involving 18 persons. The corpus will be made publicly available. Initial analyses indicate that the *single* debater speaks more often, and is looked at more often, also when corrected for the amount of speaking time. Moreover, it was found that the *single* debater is more likely to speak after one of the *team* debaters.

In future work, we will look more closely at the different speaking patterns, especially in combination with the head movements of the participants. We will also analyze whether we can automatically predict debate outcome and perceived dominance and team cooperation from the recorded data. The corpus also can be extended to be annotated in relation to the linguistic content of the phrases or with semantic connotations, in order to study their relation between head movements, in particular in a debating setting.

6. Acknowledgments

This research has been supported by the EU's FP7 Program SSPNet. The first author also would like to thank the Directorate General for Higher Education, Indonesian Ministry of Education and Culture (DIKTI) for the scholarship that allows him to take the PhD program at the University of Twente.

7. References

Aubrey, A. J., Marshall, D., Rosin, P. L., Vendeventer, J., Cunningham, D. W., & Wallraven, C. (2013). Cardiff conversation database (CCDb): A database of natural dyadic conversations. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 277–282.

Battersby, S., & Healey, P. (2010). Using head movement to detect listener responses during multi-party dialogue. In *Proceedings of the International Conference on Language Resources and Evaluation Workshop on Multi-Modal Corpora*, pp. 11–15.

Bohus, D., & Horvitz, E. (2011a). Decisions about turns in multiparty conversation: from perception to action. In *Proceedings of the 13th International Conference on Multimodal Interfaces,* pp. 153–160.

Bohus, D., & Horvitz, E. (2011b). Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Proceedings of the SIGDIAL,* pp. 98–109. Association for Computational Linguistics Stroudsburg, PA, USA.

Bousmalis, K., Mehu, M., & Pantic, M. (2013). Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image and Vision Computing*, *31*(2), pp. 203–221.

Brugman, H., & Russel, A. (2004). Annotating Multi-media/Multi-modal Resources with ELAN. In *Proceedings of the 4th International Conference on Language Resources and Evaluation,* pp. 2065–2068.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, *41*(2), pp. 181–190.

Cerrato, L., & Skhiri, M. (2003). Analysis and measurement of head movements signalling feedback in face-to-face human dialogues. In *Proceedings of First Nordic Symposium on Multimodal Communicati on,* pp. 43–52.

Curhan, J. R., & Pentland, A. (2007). Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *The Journal of Applied Psychology*, *92*(3), pp. 802–11.

De Kok, I., Ozkan, D., Heylen, D., & Morency, L.-P. (2010). Learning and evaluating response prediction models using parallel listener consensus. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction,* pp. 1.

Duncan Jr, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*, pp. 283–292.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), pp. 504–528.

Hadar, U., Steiner, T. J., & Clifford Rose, F. (1985). Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, *9*(4), pp. 214–228.

Heylen, D. (2005). Challenges ahead: head movements and other social acts during conversations. In *Proceedings of the Joint Symposium on Virtual Social Agents*, pp. 45–52.

Heylen, D. (2006). Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics*, *3*(3), pp. 241–267.

Ishi, C. T., Ishiguro, H., & Hagita, N. (2014). Analysis of relationship between head motion events and speech in dialogue conversations. *Speech Communication*, *75*, pp. 233-243.

Jovanovic, N., & Akker, R. op den. (2004). Towards automatic addressee identification in multi-party dialogues. In *Proceedings of the 5th SIGDIAL Workshop on Discourse and Dialogue*, pp. 89–92.

Pesarin, A., Cristani, M., Murino, V., & Vinciarelli, A. (2012). Conversation analysis at work: Detection of conflict in competitive discussions through semi-automatic turn-organization analysis. *Cognitive Processing*, *13 Suppl 2,* pp. 533–40.

Salamin, H., & Vinciarelli, A. (2012). Automatic role recognition in multiparty conversations: An approach based on turn organization, prosody, and conditional random fields. *IEEE Transactions on Multimedia*, *14*(2), pp. 338–345.

Sun, X., Lichtenauer, J., & Valstar, M. (2011). A multimodal database for mimicry analysis. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction,* pp. 367–37).

Verbree, D., Rienks, R., & Heylen, D. (2006). First steps towards the automatic construction of argument diagrams from real discussions. In *Proceedings of the Conference on Computational Models of Argument*, pp. 183–194.

Vertegaal, R., Slagter, R., Veer, G. van der, & Nijholt, A. (2001). Eye Gaze Patterns in Conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* pp. 301–308.

Vinciarelli, A., Dielmann, A., Favre, S., & Salamin, H. (2009). Canal9: A database of political debates for analysis of social interactions. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction Workshops,* pp. 1–4.

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., & Schroeder, M. (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, *3*(1), pp. 69–87.