Towards a Methodology Supporting Semiautomatic Annotation of Head Movements in Video-recorded Conversations

Patrizia Paggio
University of Copenhagen
University of Malta
paggio@hum.ku.dk

Costanza Navarretta
University of Copenhagen
costanza@hum.ku.dk

Bart Jongejan
University of Copenhagen
bartj@hum.ku.dk

Manex Agirrezabal
University of Copenhagen
manex.aguirrezabal@hum.ku.dk

Abstract

We present a method to support the annotation of head movements in video-recorded conversations. Head movement segments from annotated multimodal data are used to train a model to detect head movements in unseen data. The resulting predicted movement sequences are uploaded to the ANVIL tool for post-annotation editing. The automatically identified head movements and the original annotations are compared to assess the overlap between the two. This analysis showed that movement onsets were more easily detected than offsets, and pointed at a number of patterns in the mismatches between original annotations and model predictions that could be dealt with in general terms in post-annotation guidelines.

1 Introduction

Communicative gestures, such as head movements, gaze and hand gestures, are important in face-to-face communication, where they provide relevant and non redundant information that must be processed together with speech (Kendon, 2004). The automatic annotation of gestures is still problematic while their manual annotation is challenging and resource intensive.

Head movements are the most frequent gesture type and their importance as feedback and turn management signals has been recognised in numerous studies (Duncan, 1972; Hadar et al., 1983; McClave, 2000; Heylen et al., 2007; Allwood et al., 2007). For this reason, the automatic identification of the most frequent types of head movement has often been addressed as a component in multimodal systems where humans communicate with software agents or robots (Morency, 2009; Germesin and Wilson, 2009; Türker et al., 2018).

Many factors influence the frequency and type of gestures produced by speakers, e.g. the language (Navarretta et al., 2011) and the communicative setting. Therefore, it is important to have reliably annotated data of conversations in different languages and contexts. There are a wealth of video-recorded monologues and conversation on the internet that could be used for multimodal analysis. However, although several projects have developed multimodal data annotated with gestural information (for a discussion, see Wagner et al., 2014), few of these efforts have yet resulted in freely available annotated corpora for multiple languages such as in Koutsombogera and Vogel (2018). In parallel with this work, the linguistics community interested in co-speech gesture has created and annotated many multimodal data collections, but typically for experimental purposes and thus in very specific contexts (Holler, 2013).

In the NLP community there is a growing interest for multimodal sentiment analysis, in which visual and acoustic features are used to analyse sentiment and emotion in video data (Zadeh et al., 2017; Soleymani et al., 2017). The data collections used for this task are consequently annotated with sentiment and emotion labels, not with gestural information. A recent example is the CMU-MOSEI corpus (Zadeh et al., 2018).

This paper describes work aimed to support the annotation of head movements in video-recorded spontaneous conversations. More specifically, we propose a method in which automatically identified head movement segments from dyadic conversations are uploaded to the ANVIL multimodal annotation tool (Kipp, 2004), from which human annotators will be able to correct them. To assess how demanding it would be for human annotators to work with the system's suggestions, we compare the automatically identified head movements with manually annotated ones from the same corpus, and present both quantitative and qualitative analyses of this comparison. Finally we discuss possible improvements of the system and future work.

In section 2 we present background and related studies; in section 3 we describe the corpus, particularly the head movement annotations; in section 4 we give a succinct account of the method used to train models that predict head movements; in section 5 we analyse the results in terms of how well they match what the human annotators did; finally, in section 6 we summarise our results and suggest some feature perspectives.

2 Related Studies

The automatic identification of head movements can be based on the use of different tracking systems, e.g. looking at the coordinates provided by Kinect (Wei et al., 2013) or by various eye-tracking systems (Al-Rahayfeh and Faezipour, 2013). A number of researchers have worked with automatic detection of head movements in video-recordings of human-human and human-robot conversations, and the results of these studies have been evaluated using the manual annotations of the relevant movements. Two main approaches have been used. In the first one, computer vision techniques have been applied (Murphy-Chutorian and Trivedi, 2009; Gavrila, 1999), but most of these attempts have posed certain requirements to the quality, light and settings of the videos. The second approach concerns training various classifiers on automatically extracted visual and, possibly, audio features. For the visual features the two freely available systems OpenCV[1] and OpenPose[2] have been used, while audio features have been extracted via different systems. The focus of the majority of these studies has been that of determining the best visual and auditory features and the most effective classifiers for the task (Morency et al., 2005, 2007; Morency, 2009; Germesin and Wilson, 2009; Jongejan et al., 2017; Frid et al., 2017; Ambrazaitis and House, 2017; Paggio et al., 2020).

Only a couple of studies have addressed the integration of automatically identified head movements in annotation tools. Jongejan (2012) integrated the OpenCV facetracker into ANVIL to support head movement annotation for users of that system. The Max Planck Institute for Psycholinguistics and the Fraunhofer institutes HHI and IAIS developed AVAtech (Lenkiewicz et al., 2012), a gesture recogniser which is integrated into the ELAN multimodal annotation system (Wittenburg et al., 2006). The recogniser identifies head movements and hand gestures in videos. The system is based on skin recognition and other computer vision techniques, and therefore only works when videos follow specific requirements with respect to video quality, background colour, position of the speaker(s) and light. The system works best in cases when only one speaker is recorded and both face and hands can clearly been seen. When successful, AVAtech generates a video where the gestures are marked, and ELAN tracks with the head and gesture annotations given certain specifications. The system is not being further-developed, but it is still an integrated part of ELAN. We tested the recogniser on our data, but it failed to process the videos because of their quality, and in some cases, their format.

Given this background, additional efforts are needed to provide the research community with more and better automatic support for the creation of gestural annotation of conversational data.

3 The corpus

The corpus used in our study is the Danish NOMCO corpus of first encounter dialogues (Paggio and Navarretta, 2016), which consists of twelve spontaneous conversations between young people (six males and six females) of about five minutes each, for a total recording time of approximately one hour. Each participant took part in two dialogues with persons of different genders. During the encounters, the participants were standing facing each other and were recorded by three cameras. For this study we used recordings in which two frontal views of the participants are joined as shown in Figure 1.

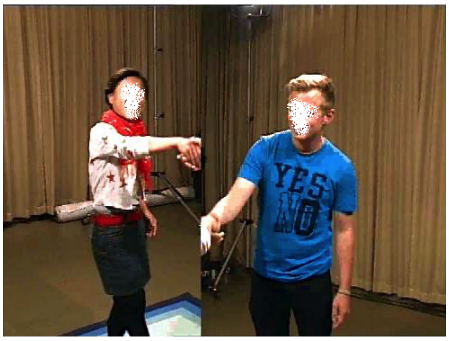

Figure 1: Screenshot from one of the NOMCO dialogues: split view

The videos were transcribed and annotated in many different ways using the ANVIL tool, and following the guidelines provided by the MUMIN

[1] https://opencv.org/

[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose

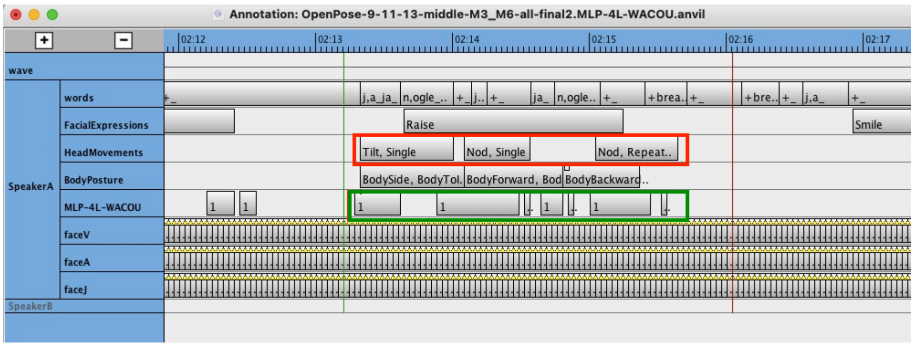

Figure 2: Screenshot of the ANVIL annotation board with manually and automatically annotated head movement sequences outlined in red and green, respectively

coding scheme for formal and functional annotation of gestural behaviour (Allwood et al., 2007). This study is concerned with the formal annotation of the speakers' head movements, for which the coding scheme provides the following labels: *Nod, Up-nod, Shake, HeadBackward, HeadForward, SideTurn, Tilt, Waggle* and *HeadOther*. A head movement is annotated by defining start and end point of the movement on the tool's annotation board, and by picking one of the labels.

The linguistic context was taken into account in order to select and annotate only those head movements that were judged to have a communicative function. To familiarise themselves with the annotation scheme and guidelines, the three annotators involved went through an initial training exercise in which they annotated the gestures in one video together, and discussed all disagreements with one another and with two expert researchers. After that, they annotated a second video independently and their mutual agreement was measured.

The inter-coder agreement for identification and classification of head movements reached a Cohen's *kappa* score in the range 0.72-0.8 (Navarretta et al., 2011). To produce the final version of the annotations, all disagreements were adjudicated by involving an expert annotator.

The annotations have been used in numerous studies addressing aspects such as the identification of feedback and turn-taking signals given through head movements and facial expressions (Paggio and Navarretta, 2011; Navarretta and Paggio, 2013). The manual annotations of head movements, as well as of other gesture types, are available for research[3].

4 Automatic detection of head movement

In this section we summarise the method we use for the automatic detection of head movements. A more detailed explanation and discussion is provided in Paggio et al. (2020).

The manual annotations of head movements were used as a basis for training a classifier to detect the presence of head movement in a frame-wise fashion. To create a speaker-independent model, the classifier had to detect movement for each speaker after having being trained on the data from the other eleven participants following a leave-one-out cross validation strategy.

The training data consisted of visual, acoustic and textual features. OpenPose (Cao et al., 2018) was used to extract nose tip positions from the data. For each of these positions we include both the cartesian (x and y) and polar (radius and angle) coordinates, thus, four different scalar values. We then calculate velocity, acceleration and jerk of these positions by computing the first, second and third order derivatives, using the previous 9, 11 and 13 frames, respectively. We include these three orders of derivatives, resulting in a total of twelve visual features.

Intensity and pitch measurements were extracted from the audio files using PRAAT (Boersma and Weenink, 2009). Finally, head movement labels, as well as word information, were extracted from the annotations.

As a result, for each frame in each video a vector was created with labels expressing presence/absence of movement and the head movement class; velocity, acceleration and jerk features; pitch and intensity values referring to the gesturer; and finally a binary feature expressing whether the same gesturer is speaking or not.

Several classifiers were tested on the data. The

[3] https://cst.ku.dk/projekter/the-danish-nomco-corpus/

best performing one was a Multilayer Perceptron (MLP) with 4 layers, which achieved an average accuracy across the twelve speakers of 0.730 and an F1 score of 0.684 (macro average).

The frame-wise head movement predictions were combined into sequences by conjoining directly adjacent frames for which a positive movement value had been assigned by the predictor and mapped onto the XML format necessary to read them into the ANVIL tool as independent tracks. In this way, a comparison could be carried out between the manual and the automatic annotations.

Figure 2 is a screenshot of the ANVIL annotation board visualising an excerpt of the annotation with tracks for one of the speakers (Speaker A). There are eight tracks corresponding to i) the speech transcription, ii-iv) the manual annotations of facial expressions, head movements and body posture, v) the head movement annotations detected by the MLP classifier, and vi-viii) the values for movement velocity, acceleration and jerk. The two tracks of interest here are the ones containing the manually annotated head movements, a sequence of which has been highlighted in red, and the corresponding automatically derived head movement elements, marked in green. In this example, we see that three head movements in the former track roughly correspond to three larger and two smaller elements in the latter. There are also, however, two additional predictions in the middle of the sequence which do not overlap with any annotated movement. In a scenario in which a new annotation task has to be produced, only the elements suggested by the model would be available, making it possible for an annotator to revise these suggestions as needed.

To get an impression of how much and what type of revision may become necessary, in the next section we look more closely at how well the two annotation tracks correspond to one other.

5 Analysis of results

5.1 Overlaps between manual and automatic annotations

The total number of annotated head movements in the Danish NOMCO corpus is 3117. The number of movement sequences predicted by the MLP binary classifier, however, is 7827.

The mean duration of an annotated movement sequence is 935 ms (sd: 579) while the mean duration of a predicted movement is 286 ms (sd: 370). Thus the model predicts a much higher number of shorter movements compared to the movement sequences identified by the human annotators. The duration distribution of annotated vs predicted movement is visualised in a combined boxplot in Figure 3. It is evident that the predicted movements are much shorter on average, however there are also large numbers of outliers in the upper end of both plots.

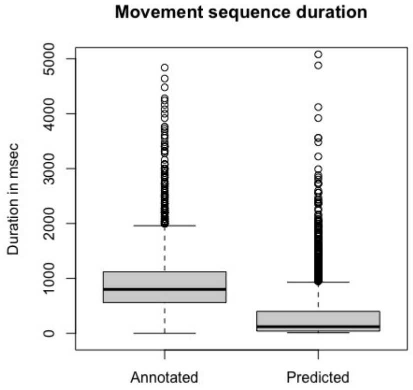

Figure 3: Distribution of the duration of head movements in the manual annotations vs model predictions

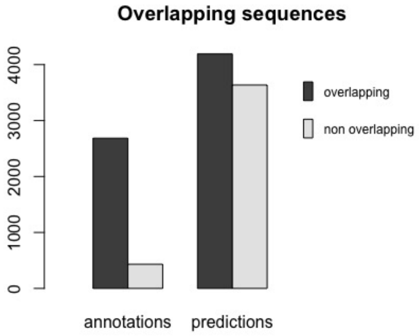

Figure 4: Number of overlaps in annotated and predicted head movements

To reach a general understanding of the way the two sets of movements relate to one another in terms of alignment, overlaps between annotated and predicted movement sequences were extracted. There are multiple overlaps between the annotated and the predicted movement sequences, such that 2685 of the annotated sequences (86%) overlap with one or more predicted ones. Conversely, 4191 of the predicted sequences (53%) overlap with an annotated movement. In other words, the model has quite good recall, but also generates a relatively high number of false positives (predictions that do

Overlaps per annotation	Annotated elements #	Annotated elements %	Total predicted elements
One	1234	46	1234
Two	751	28	1502
Three	385	14	836
Four	209	8	285
Five	57	2	126
Six-Nine	49	2	208
Total	2685	100	4191

Table 1: Breakdown of number of overlaps per movement annotation

not overlap with manual annotation). A visualisation is provided in Figure 4.

Table 1 shows a breakdown of the overlaps. The table summarises how many annotated movements overlap with one predicted movement, two of them, three and so on. As can be seen, in the vast majority of cases (88%), an annotated movement overlaps with 1-3 predicted movement sequences. The example displayed in Figure 2 shows one such situation, in which three annotated elements correspond to five predicted ones. In 12% of the cases one annotation corresponds to more than three predictions, and marginally (only two cases in fact) to as many as nine different predicted elements.

For each annotated head movement that overlapped with some predicted movement, the longest overlap of at least 0.01 ms was then extracted. This generated a list of 2661 overlaps. The average duration of the longest overlaps is 445.83 ms (sd: 317.06).

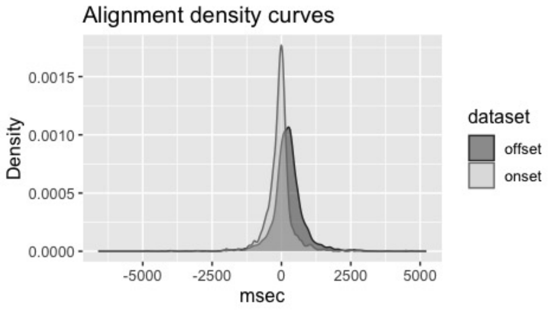

Figure 5: Overlap alignment at onset and offset points

None of these overlaps are complete. In other words, in no case does an annotated movement align completely both at onset and offset with a predicted one. To investigate whether the longest overlaps predicted by the model align better at onset or offset, we calculated the differences at start and end points between the annotated and the predicted movement for each overlap. The result is two lists of misalignments, which are visualised by means of a combined density plot in Figure 5.

The onset misalignment is -68 ms on average (sd: 500), indicating that the overlapping predicted sequence tends to start slightly after the annotated one (start of annotated − start of predicted). In comparison, the mean offset misalignment is 232 ms (sd: 613), indicating that the offset of the overlapping predicted sequence tends to occur earlier than the offset of the corresponding annotated movement (end of annotated − end of predicted). In other words, annotated and predicted movements align better at onset than at offset, albeit with considerable variation. The difference is significant on a Wilcoxon rank sum test (W = 2019582, p-value < 2.2e-16).

Here we advance some conjectures for why onsets may be easier for the model to detect. Predicting the offset of a movement may be more difficult due to inherent characteristics of head movements, and perhaps gestural movements in general, with movements petering out in a way that makes it difficult to distinguish the exact end point. An additional difficulty may also be due to movements containing holds, where the human annotators would consider such a hold part of a movement which has not yet come to an end, whereas automatic detection would probably divide such a movement up in several sequences. In fact in a number of cases, the offset misalignment is due to the fact that multiple sequences overlap with the tail of the annotated movements.

5.2 Comparing manual and automatic annotations: a qualitative analysis

In this section, we present a qualitative comparison of the manual and automatic annotations in ANVIL, focusing on the annotations of the head movements of one participant in a randomly chosen video. We distinguish four different situations in which there are differences between an annotated movement sequence and what should be the corresponding predicted one.

We start with cases in which head movements identified by the model were not annotated by the coders. There are several explanations for this difference between the two types of annotations:

a. The participant moves the head, but the gesture has not been annotated because it does not have a clear communicative function, e.g. the participant moves the head to look at the other speaker immediately before they shake hands at the beginning of the interaction.

b. The participant moves the whole body or her shoulders and the head moves with the rest of the body/upper body. In such cases, the annotators were instructed to code the movement as a body posture shift.

c. A non-movement is identified as a movement by the system (true error).

The discrepancy explained under (a) is due to the annotation guidelines adopted in the NOMCO project (op. cit.), in which a basic distinction between communicative and non-communicative gestures was adopted. An automatic annotator could not be expected to make such a distinction, and in fact other annotation projects may take a different approach to this issue. Cases subsumed under (b), in contrast, could perhaps be avoided with a more careful use of visual features in which head and body movement are handled separately. At the moment, in fact, we are not using any visual body markers from OpenPose. As for false positives due to true errors as in (c), they can probably not be entirely avoided. More analysis would be necessary to know exactly how many of these cases we have.

There are then a number of cases in which one manually annotated head movement overlaps with several automatically identified ones, again due to several reasons:

d. The manually identified head movement is a repeated gesture, e.g. a repeated nod was coded as one uninterrupted gesture whereas the system found a sequence of smaller independent movements. However, it must be noted that in some cases a manually coded repeated head movement is also identified as a single head movement by the system.

e. A movement of the upper body partially overlaps with a head movement in the manual annotations, whereas the model found two separate head movements.

f. The annotated movement contains one or several holds, or subsequences in which the movement slows down to then pick up in velocity again. The model typically identifies a number of separate movement sequences in such cases.

Cases such as described under (d) are relatively well-defined and could be handled in future annotations by instructing the coders to conflate the automatically generated components of a repeated gesture into one uninterrupted element. Cases under (e) are similar to (b) above. The same comments therefore hold. The situation described under (f), which was also discussed in the preceding section, requires some consideration. An annotator would have to judge whether the distinct elements should be considered part of an uninterrupted gesture and therefore joined, or not. A typical example of this are tilts of the head. Quite often, the speaker who performs a head tilt will keep the tilt for a few seconds – sometimes longer than that – before bringing the head back to normal position. In NOMCO a decision was made always to annotate such cases as uninterrupted gestures, however the hold may sometimes be so long as to warrant a different solution.

The third mismatch type considered consists of cases in which one automatically identified head movement corresponds to several manually annotated head movements. All of these seem in fact explainable in the same way:

g. The manually identified head movements are different but consecutive ones, e.g. a manually identified head forward movement followed by a repeated shake overlap with a single predicted movement.

Since the model is a binary classifier, no other solutions would have been possible for the (g) cases. Two scenarios can be envisaged here. In the first one, an annotator can be instructed to split a movement element into a sequence of separate ones if it corresponds to several different gestures. In the second one, a more complex model must be trained to detect not just presence of head movement but different movement classes.

Finally, there are cases in which the annotators coded a head movement that is clearly visible in the video, but which the system failed to identify. No explanation could be provided for such cases, which are, however, relatively rare. Recall in fact that 86% of the annotated head movements overlap with at least one predicted one.

6 Conclusion and future work

In this paper we have presented work aimed at creating support for the annotation of head movements in video-recorded data, an annotation which is a necessary first step in any analysis of the role played by head movements in the semantics and pragmatics of face-to-face conversations.

The first step in our methodology is to train machine learning models to detect head movements based on visual and acoustic features that are obtained from video-recorded materials without specific demands on the audio-visual quality. The second step is to make the automatically derived movement sequences available in the ANVIL multimodal annotation tool. The evaluation of the results showed a good coverage in that 86% of annotated movements were predicted with various degrees of overlap. However, a high number of non-existing movements were also incorrectly predicted.

Subsequent analysis showed that movement onsets were more easily detected than offsets, and pointed at a number of patterns in the mismatches between original annotations and model predictions that could be dealt with in general terms in post-annotation guidelines.

More work is needed to understand and possibly reduce the impact of false positives. The sequence-generation process may also be a source of errors, thus alternative methods to combine frames into sequences should be explored. Another obvious extension of this work concerns the automatic classification of head movement in different types. Moreover, we plan to work on adapting our results to make the predicted movement sequences compatible with the ELAN tool.

Finally, we need to determine how useful the automatic prediction is in a concrete annotation exercise. It is important to evaluate not only whether the annotators would save time post-editing the predicted movements rather than creating new ones from scratch, but also whether the same accuracy can be reached and what kinds of error the annotators would make in the post-editing scenario compared to the traditional one. We plan to conduct such a study in connection with the development of a new corpus of online group discussions that is being collected by the international network on GEstures and Head Movements in language (GEHM)[4]

Acknowledgements

This work was supported by the Independent Research Fund Denmark (grant 9055-00004B). We would also like to thank the anonymous reviewers for their useful comments.

[4] https://cst.ku.dk/english/projects/gestures-and-head-movements-in-language-gehm/

References

Amer Al-Rahayfeh and Miad Faezipour. 2013. Eye tracking and head movement detection: A state-of-art survey. *IEEE Journal of Translational Engineering in Health and Medicine*, 1:2100212 –2100212.

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In Jean-Claude Martin, Patrizia Paggio, Peter Kuehnlein, Rainer Stiefelhagen, and Fabio Pianesi, editors, *Multimodal Corpora for Modelling Human Multimodal Behaviour*, volume 41 of *Special issue of the International Journal of Language Resources and Evaluation*, pages 273–287. Springer.

Gilbert Ambrazaitis and David House. 2017. Acoustic features of multimodal prominences: Do visual beat gestures affect verbal pitch accent realization? In *Proceedings of The 14th International Conference on Auditory-Visual Speech Processing (AVSP2017)*, pages 89–94. KTH.

Paul Boersma and David Weenink. 2009. Praat: doing phonetics by computer (version 5.1.05) [computer program]. Retrieved May 1, 2009, from http://www.praat.org/.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multiperson 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292.

Johan Frid, Gilbert Ambrazaitis, Malin Svensson-Lundmark, and David House. 2017. Towards classification of head movements in audiovisual recordings of read news. In *Proceedings of the 4th European and 7th Nordic Symposium on Multimodal Communication (MMSYM 2016)*, 141, pages 4–9, Copenhagen. Linköping University Electronic Press, Linköpings universitet.

Dariu M Gavrila. 1999. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82 – 98.

Sebastian Germesin and Theresa Wilson. 2009. Agreement detection in multiparty conversation. In *Proceedings of ICMI-MLMI 2009*, pages 7–14.

Uri Hadar, Timothy J Steiner, Ewan C Grant, and Frank Clifford Rose. 1983. Head Movement Correlates of Juncture and Stress at Sentence Level. *Language and Speech*, 26(2):117–129.

Dirk Heylen, Elisabetta Bevacqua, Marion Tellier, and Catherine Pelachaud. 2007. Searching for prototypical facial feedback signals. In *Proceedings of 7th International Conference on Intelligent Virtual Agents*, pages 147–153.

Judith Holler. 2013. *52. Experimental methods in co-speech gesture research*, pages 837–857. De Gruyter Mouton.

Bart Jongejan. 2012. Automatic annotation of head velocity and acceleration in anvil. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 201–208.

Bart Jongejan, Patrizia Paggio, and Costanza Navarretta. 2017. Classifying head movements in video-recorded conversations based on movement velocity, acceleration and jerk. In *Proceedings of the 4th European and 7th Nordic Symposium on Multimodal Communication (MMSYM 2016), Copenhagen, 29-30 September 2016*, 141, pages 10–17. Linköping University Electronic Press, Linköpings universitet.

Adam Kendon. 2004. *Gesture*. Cambridge University Press.

Michael Kipp. 2004. *Gesture Generation by Imitation – From Human Behavior to Computer Character Animation*. Boca Raton, Florida: Dissertation.com.

Maria Koutsombogera and Carl Vogel. 2018. Modeling collaborative multimodal behavior in group dialogues: The multisimo corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Przemyslaw Lenkiewicz, Binyam Gebrekidan Gebre, Oliver Schreer, Stefano Masneri, Daniel Schneider, and Sebastian Tschöpel. 2012. Avatech — automated annotation through audio and video analysis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Evelyn McClave. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878.

Louis-Philippe Morency. 2009. Co-occurrence graphs: contextual representation for head gesture recognition during multi-party interactions. In *Proceedings of the Workshop on Use of Context in Vision Processing*, pages 1–6.

Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. 2007. Latent-dynamic discriminative models for continuous gesture recognition. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

Louis-Philippe Morency, Candace Sidner, Christopher Lee, and Trevor Darrell. 2005. Contextual recognition of head gestures. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 18–24.

Erik Murphy-Chutorian and Mohan Manubhai Trivedi. 2009. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626.

Costanza Navarretta, Elisabeth Ahlsén, Jens Allwood, Kristiina Jokinen, and Patrizia Paggio. 2011. Creating Comparable Multimodal Corpora for Nordic Languages. In *Proceedings of the 18th Conference Nordic Conference of Computational Linguistics*, pages 153–160, Riga, Latvia.

Costanza Navarretta and Patrizia Paggio. 2013. Classifying Multimodal Turn Management in Danish Dyadic First Encounters. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (Nodalida 2013)*, pages 133–146, Oslo, Norway. NEALT.

Patrizia Paggio, Manex Agirrezabal, Bart Jongejan, and Costanza Navarretta. 2020. Automatic detection and classification of head movements in face-to-face conversations. In *Proceedings of LREC2020 Workshop "People in language, vision and the mind" (ONION2020)*, pages 15–21, Marseille, France. European Language Resources Association (ELRA).

Patrizia Paggio and Costanza Navarretta. 2011. Head movements, facial expressions and feedback in Danish first encounters interactions: A culture-specific analysis. In *Universal Access in Human-Computer Interaction - Users Diversity. 6th International Conference. UAHCI 2011, Held as Part of HCI International 2011*, number 6766 in LNCS, pages 583–690, Orlando Florida. Springer Verlag.

Patrizia Paggio and Costanza Navarretta. 2016. The Danish NOMCO corpus: Multimodal interaction in first acquaintance conversations. *Language Resources and Evaluation*, pages 1–32.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Bekir Berker Türker, E. Erzin, Y. Yemez, and T. M. Sezgin. 2018. Audio-visual prediction of head-nod and turn-taking events in dyadic interactions. In *INTERSPEECH*, pages 1741–1745.

Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209 – 232.

Haolin Wei, Patricia Scanlon, Yingbo Li, David S Monaghan, and Noel E O'Connor. 2013. Real-time head nod and shake detection for continuous human affect recognition. In *14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4. IEEE.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, pages 1556–1559.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.