Proceedings of ONION 2020: Workshop on peOple in laNguage, vIsiOn and the miNd, pages 15–21
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Automatic Detection and Classification of Head Movements in Face-to-Face Conversations

Patrizia Paggio[1,2], **Manex Agirrezabal**[1], **Bart Jongejan**[1], **Costanza Navarretta**[1]
[1]University of Copenhagen, [2]University of Malta
paggio@hum.ku.dk, manex.aguirrezabal@hum.ku.dk, bartj@hum.ku.dk, costanza@hum.ku.dk

Abstract

This paper presents an approach to automatic head movement detection and classification in data from a corpus of video-recorded face-to-face conversations in Danish involving 12 different speakers. A number of classifiers were trained with different combinations of visual, acoustic and word features and tested in a leave-one-out cross validation scenario. The visual movement features were extracted from the raw video data using OpenPose, the acoustic ones from the sound files using Praat, and the word features from the transcriptions. The best results were obtained by a Multilayer Perceptron classifier, which reached an average 0.68 F1 score across the 12 speakers for head movement detection, and 0.40 for head movement classification given four different classes. In both cases, the classifier outperformed a simple most frequent class baseline, a more advanced baseline only relying on velocity features, and linear classifiers using different combinations of features.

Keywords: head movement detection, multimodal corpora, visual and speech features

1. Introduction

Head movements play an important role in face-to-face communication in that they provide an effective means to express and elicit feedback, and consequently establish grounding and rapport between speakers; they contribute to turn exchange; they are used by speakers to manage their own communicative behaviour, e.g. in connection with lexical search (Allwood, 1988; Yngve, 1970; Duncan, 1972; McClave, 2000). Therefore, it is crucial for conversational systems to be able to identify and interpret speakers' head movements as well as generate them correctly when interacting with users (Ruttkay and Pelachaud, 2006).

This paper is a contribution to the automatic identification of head movements from raw video data coming from face-to-face dyadic conversations. It builds on previous work where a number of models were trained to detect head movements based on movement and speech features, and extends that work in several directions by extracting movement features using newer software, by trying to distinguish between different kinds of movement, and by training and testing speaker-independent models based on a larger dataset.

The paper is structured as follows. In section 2 we discuss related work in the area. Section 3 is dedicated to the features for the prediction of head movements. In section 4 we present the corpus that we used for the current study. Finally in section 5 we discuss the results and propose some possible future directions.

2. Related work

Several studies have been relatively successful in performing head movement detection from tracked data, for example by using coordinates obtained through eye-tracking (Kapoor and Picard, 2001; Tan and Rong, 2003) or Kinect sensors (Wei et al., 2013). A different approach to the task is to detect head movements in raw video data. Such an approach has the potential of making available large amount of data to train systems to deal with multimodal communication in different languages and communicative scenarios. Large annotated multimodal corpora are in turn a prerequisite to the development of natural multimodal interactive systems. Surveys of the way computer vision techniques can be applied to gesture recognition are given in Wu and Huang (1999) and Gavrila (1999). Both works conclude, however, that the field is still a fairly new one, and many problems remain as yet unsolved.

Work has also been done trying to detect gestures based on visual as well as language or speech features. In this line of research, Morency et al. (2005) proposed a methodology where SVM and HMM models were trained to predict feedback nods and shakes in human-robot interactions. The visual features used for head movement recognition were enriched with features from the dialogue context. It can be argued, however, that human-robot interaction is much more constrained than spontaneous human dialogue, and thus the task of predicting the user's head movements is probably easier, or at least different than in human-human communication data. In Morency et al. (2007), models were trained to recognise head movements in video frames in a variety of datasets based on visual features obtained from tracked head velocities or eye gaze estimates extracted from video data. A number of different models were compared in the study, and it was found that LDCRF (Latent-Dynamic Conditional Random Field) was the best performing of the models. The authors attribute the result to the fact that the model is good at dealing with unsegmented sequences, in this case movement sequences. Morency (2009) studied the co-occurrence between head gestures and speech cues such as specific words and pauses in multi-party conversations, and relevant contextual cues were used to improve a vision-based LDCRF head gesture recognition model.

In Jongejan (2012), OpenCV was applied to the detection of head movement from videos based on velocity and acceleration, in combination with customisable thresholds, for the automatic annotation of head movements using the ANVIL tool (Kipp, 2004). The obtained annotations correlated well with the manual annotation at the onset, but generated a high number of false positives. In Jongejan et al. (2017),

three visual movement features were used to train an SVM classifier of head movement.

Frid et al. (2017) used the corpus of read news in Swedish described in Ambrazaitis and House (2017) to detect head movements that co-occur with words. The head movements were manually annotated and OpenCV for frontal face detection was used in order to calculate velocity and acceleration features. A Xgboost classifier was trained to predict absence or presence of head movements co-occurring with words.

Acoustic features have also been used for head movement prediction. For example Germesin and Wilson (2009) combined pitch and energy of voice with word, pause and head pose information to identify agreement and disagreement signals in meeting data. Such work is based on linguistic and psycho-linguistic findings that have shown a tight relationship between facial movements and acoustic prominence, to the point of talking about audiovisual prominence (Granström and House, 2005; Swerts and Krahmer, 2008; Ambrazaitis and House, 2017).

In the work by Paggio et al. (2018), movement features were considered together with acoustic features to identify head movements in conversational data. The authors performed several experiments with different feature sets and also, several prediction paradigms were tested, including common classifiers and sequence-based models. It was observed that a Multilayer Perceptron showed the best results when trained on one speaker and tested on another one. In this study, we build on those preliminary results by extending our dataset to consider twelve different speakers, and we experiment with the classification of different head movement types.

3. Predictive features

Similarly to what was done in Paggio et al. (2018), three time-related derivatives with respect to the changing position of the head are used here as features for the identification of head movements: *velocity*, *acceleration* and *jerk*. Velocity is change of position per unit of time, acceleration is change of velocity per unit of time, and finally jerk is change of acceleration per unit of time. We suggest that a sequence of frames for which jerk has a high value either horizontally or vertically may correspond to the *stroke* of the movement (Kendon, 2004).

OpenPose (Cao et al., 2018) was used to extract nose tip positions from the data. Using a sliding window, velocity, acceleration and jerk values were computed for video frame sequences using a polynomial (linear, quadratic and cubic, respectively) regression over a number of observations of nose tip positions. Several window frames were experimented with. The results reported in this paper were obtained by considering 9 frames for velocity, 11 for acceleration and 13 for jerk. For each of the three derivatives, four values are computed for each frame and used to train the models. The 12 values are both the cartesian (x and y) and polar (radius and angle) coordinates of the velocity, acceleration and jerk vectors. Since we analyse video data, we do not have depth information, and so we are restricted to express velocity, acceleration and jerk as vectors in a two dimensional plane. Angle values have integer values between 1 and 12, like the directions on a clock dial.

It must be noted that the video recordings are characterised by 25 frames per second and a resolution of either 640x360 (.avi) or 640x369 (.mov). Thus the quality is quite low given today's standards. In addition, since the participants are recorded almost in full height, the head movements are very tiny when expressed in pixels. All of this is bound to have an effect on how accurately the movement derivatives can predict head movement.

Acoustic features were extracted from the speech channels of all speakers using the PRAAT software (Boersma and Weenink, 2009). In general, several studies indicate that head movements are likely to occur together with prosodic stress, whereas the opposite is not necessarily true (Hadar et al., 1983; Loehr, 2007). Since in Danish, which is the language of our study, stress is expressed through fundamental frequency, vowel duration and quality, as well as intensity (Thorsen, 1980), we decided to rely on pitch and intensity features to model a possible relation between focal patterns and head movements. F0 values and intensity values were sampled with 25 frames per second as is done for the movement features and added to the training data. The hypothesis is that changes in pitch or peaks of intensity might be associated with head movement strokes, and thus help in identifying movement.

Based on the analysis of co-occurrence patterns between head movements and verbalisation in the corpus data (Paggio et al., 2017), we finally added to the predictive features information as to whether the person performing the movement, the *gesturer*, is speaking or not. This binary feature was added to each frame based on the speech transcription, which was done manually and includes word boundaries.

4. Data, training and test setup

Figure 1: Screen shot from one of the video recordings showing combined almost frontal camera views

The data used for this study is taken from the Danish NOMCO corpus (Paggio et al., 2010), a collection of twelve video-recorded first encounter conversations between pairs of speakers (half females, half males) for a total interaction of approximately one hour. Each speaker took part in two different conversations, one with a male and one with a female. The speakers are standing in front of each other. The conversations were recorded in a studio using three different cameras and two cardioid microphones. For the work presented here we used a version of the recordings in which both speakers are being viewed almost frontally, and the two views are combined in a singled video as shown in Figure 1. The data have been annotated

Movement type	No. movements	No. frames
None	NA	125,747
Nod	926	21,755
Shake	337	9,505
Other	1,854	41,053
Total movement	3,117	72,313

Table 1: Different types of head movements in the dataset: total number of frames and whole movements

	None	Nod	Shake	Other	All
Mean	10,479	1,813	792	3,421	6,026
CV	0.13	0.47	0.50	0.20	0.20

Table 2: Distribution of different head movement types in the dataset: average mean number of frames and coefficient of variation across 12 speakers

with many different annotation layers (Paggio and Navarretta, 2016), including a manually obtained speech transcription with word-specific boundaries, and temporal segments corresponding to different types of head movement (Allwood et al., 2007). The Cohen's (1960) κ score results of inter-coder agreement experiments involving two annotators are between 0.72 and 0.8 for the identification and classification of head movements (Navarretta et al., 2011). For this study, we have focused on two ways of looking at the head movements; i. distinguishing between head movement and absence of it; ii. distinguishing between nods, shakes, other kind of head movement, and no movement.

In table 1 we show the distribution of the four types of head movement in the annotated corpus both in terms of entire movement sequences and number of video frames. Thus, 3,117 head movements were annotated in total, corresponding to 72,313 movement frames. Frames containing no head movement constitute by far the majority of the video footage. The *Nod* class subsumes both down and up nods. It was singled out together with *Shake* because these two classes have been targeted previously in head movement detection studies (Morency et al., 2005). The *Other* category groups a number of distinct types in the annotation, i.e. *HeadBackward, HeadForward, SideTurn, Tilt, Waggle* and *HeadOther*.

There is of course speaker-dependent variation in the frequency of the various movement types. Table 2 displays mean averages and coefficient of variations for how different movement and non movement frames are distributed across the twelve speakers. The figures show that the frequency of occurrence of both *Nod* and *Shake* varies considerably in the speaker sample.

The duration of the head movements in the annotated corpus is 934.78 ms on average (SD: 579.44 ms). A histogram of head movement duration is given in Figure 2. Although most movements are shorter than 1500 ms, we see a long tail of outliers with a maximum duration of up to 7,080 ms.

To derive training data from the twelve annotated videos, movement, acoustic and word features were extracted as explained in the previous section so that for each frame in each video a vector was created with features expressing presence/absence of movement, a label for each of the four movement classes, four velocity, four acceleration and four jerk features, pitch and intensity values referring to the gesturer and a binary feature expressing whether the same gesturer is speaking or not.

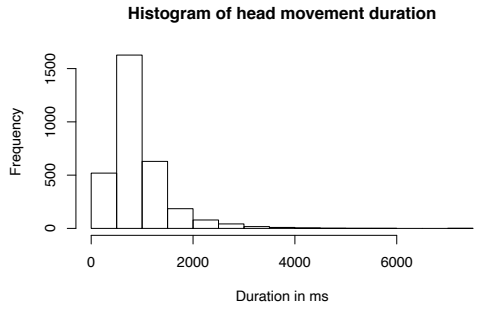

Figure 2: Duration of annotated head movements in the dataset

The data were then used to train a number of different classifiers to predict the head movements of each speaker given training data from the other eleven speakers (leave-one-out cross validation). In what follows, we will report accuracy and F1 results achieved by the various classifiers on average across speakers. It should be kept in mind, however, that there is variation across speakers in number of types of head movement produced, as already noted. Moreover, the accuracy of the classifiers may be influenced by the fact that some speakers are sometimes situated on the left and sometimes on the right, and others are in the same position in both the conversations they took part in.

As mentioned earlier, two tasks were conducted. The first is detection of head movement (irrespective of the type), and the second is classification of head movement type given the four classes *None, Nod, Shake* and *Other*.

Two baselines were chosen. The first one corresponds to the results obtained by a simple most-frequent category model, which will always predict that there is no movement in the frame. The second one is a logistic regression classifier that only uses velocity features. We then experimented with the complete range of movement derivatives (velocity, acceleration and jerk). Finally, we added acoustic and word information relative to the gesturer. The following classifier types were used to train models using the various feature combinations: i. a Logistic Regression (LR) classifier, which is an example of a simple model, ii. a linear Support Vector Machine (LINEARSVC), which was used by several earlier studies for head movement detection, and iii. a Multilayer Perceptron (MLP) with four layers, as an example of a non-linear classifier. [1]

5. Results

The results of the binary classification experiments are given in terms of average accuracy in table 3, and F1 score (macro average) in table 4. Looking at accuracy first, all models perform better than the most frequent class (MF)

[1]The data and the Jupyter notebooks that were used in our experiments can be found at https://github.com/kuhumcst/head_movement_detection.

Exp	Features	MF	LR	LINEARSVC	MLP
1	Only velocity	0.635	0.686	0.680	0.707
2	All visual features (no sound)	0.635	0.721	0.718	**0.733**
3	All visual and acoustic (only gesturer)	0.635	0.722	0.718	0.730
4	All visual and acoustic+word (only gesturer)	0.635	0.725	0.723	0.730

Table 3: Accuracy results of classification experiments (mean over 12 speakers). Classes are presence and absence of movement.

Exp	Features	MF	LR	LINEARSVC	MLP
1	Only velocity	0.387	0.575	0.557	0.648
2	All visual features (no sound)	0.387	0.644	0.633	0.684
3	All visual and acoustic (only gesturer)	0.387	0.646	0.634	0.681
4	All visual and acoustic+word (only gesturer)	0.387	0.658	0.650	**0.684**

Table 4: F1 results (macro average) of classification experiments (mean over 12 speakers). Classes are presence and absence of movement.

		Predicted as				
		None	Nod	Shake	Other	Sum
Gold value	None	**113,566**	1,984	327	9,870	125,747
	Nod	13,429	**4,528**	74	3,724	21,755
	Shake	5,977	184	**618**	2,726	9,505
	Other	23,148	2,089	584	**15,232**	41,053

Table 5: Classification of different types of head movements in the whole dataset: error matrix

Movement type	No. frames	Precision (%)	Recall (%)
None	125,747	72.74	90.31
Nod	21,755	51.54	20.81
Shake	9,505	38.55	6.5
Other	41,053	48.28	37.1

Table 6: Classification of different types of head movements in the whole dataset: total number of frames, precision and recall for each type

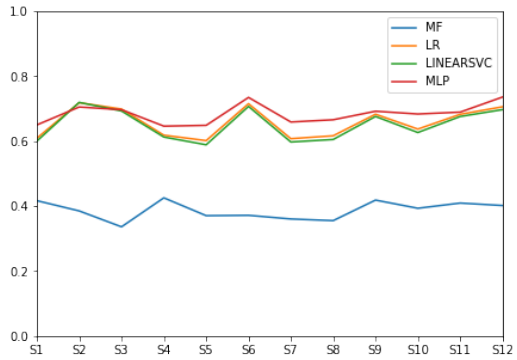

Figure 3: Visualisation of the F1-score of the binary model that include all features (exp. 4 in table 4)

baseline. We also see that the MLP classifier performs better than all the others irrespective of the combination of features used in the training. The overall best accuracy is achieved by MLP using all the three movement features, whereas acoustic and word features seem to introduce some noise (even though the difference between the MLP results in experiment 2 on the one hand and 2 and 3 on the other is marginal).

Turning to F1, we observe again that all models definitely outperform the baseline, and that the MLP classifier is consistently the best in all experiments. In this case, the best result is achieved either using the entire range of features or only the visual ones. Adding acoustic features alone produces a slightly lower F1.

Figure 3 shows how the F1 score obtained by the best binary models, i.e. those trained with the complete range of features, varies depending on the speaker. The MLP classifier is not only the best performing one on average, but also the one where the F1 score varies the least. However, there is still some variation. In fact, the standard deviation for the results achieved by MLP is 0.053 for accuracy and 0.046 for F1.

We now turn to the results of the multi-class prediction experiments, which are shown in table 7 for accuracy and table 8 for F1 score (macro average). Determining the type of head movement in a multi-class prediction scenario is a more difficult task than having to choose between movement and non-movement. Therefore, it is not surprising that the results are generally worse. Nevertheless, all the models perform better than the baseline both as regards accuracy and F1. Also in this case, MLP is generally the best classifier. If we now focus on the accuracy results first, we see again that the best accuracy is achieved by MLP when using all the movement features but no acoustic or word features. When we look at the F1 scores, however, we see that acoustic features this time not only help the classifier, but provide the best performing model in combination with movement features.

Further analysis of the results is provided by the error matrix in table 5, which relates to the best performing classifier (MLP in exp. 3). We see first of all that head movements of all types are confused with no movement, and to some extent with movements of type *Other*. Nods and shakes, on the contrary, are seldom exchanged for one another, which seems a good result given the fact that they are quite different from the point of view of their movement characteristics.

In table 6 we show precision and recall figures for the different movement types. Recall is in general low for movement frames, while precision is better. We see this as an advantage in that an automatic procedure that misses exist-

Exp	Features	MF	LR	LINEARSVC	MLP
1	Only velocity	0.635	0.648	0.646	0.657
2	All visual features (no sound)	0.635	0.660	0.657	**0.677**
3	All visual and acoustic (only gesturer)	0.635	0.661	0.658	0.676
4	All visual and acoustic+word (only gesturer)	0.635	0.668	0.665	0.679

Table 7: Accuracy results of multi-class prediction experiments (mean over 12 speakers). Classes are nod, shake, other, none.

Exp	Features	MF	LR	LINEARSVC	MLP
1	Only velocity	0.194	0.256	0.249	0.308
2	All visual features (no sound)	0.194	0.291	0.277	0.396
3	All visual and acoustic (only gesturer)	0.194	0.294	0.279	**0.397**
4	All visual and acoustic+word (only gesturer)	0.194	0.313	0.297	0.394

Table 8: F1 results (macro average) of multi-class prediction experiments (mean over 12 speakers). Classes are nod, shake, other, none.

ing head movements seems more acceptable than one that finds non-existing ones. Precision in the detection of head movements is highest for *Nod*, followed by *Other*, followed by *Shake*. The degree of precision depends not only on frequency of occurrence (there are more nods than shakes), but also on how homogeneous the classes are (the class *Other* is not as homogeneous as the class *Nod*).

6. Discussion

In general, it is difficult to compare our results directly to what other head movement detection studies have achieved because of the diversity of recording settings, number of participants, communicative situations etc. The work that resembles ours the most in terms of the methodology used is perhaps the paper by Frid et al. (2017) in that they also rely on movement derivatives. They also look at the co-occurrence of head movements and words, but do so in a different way by predicting for each word whether it is accompanied by a movement or not. Their results, 0.89 accuracy and 0.61 F1 score, are not very dissimilar from those obtained by our best model in the binary classification.

It must be noted, however, that we are detecting head movements in less favourable conditions since our subjects are recorded in full body size. In addition, the quality of our videos is, as already mentioned, not up to today's standards. Furthermore, the acoustic signal is also far from optimal because the microphones were hanging from the ceiling rather than being close to the participants' mouths.

The present study is a further development of the earlier experiment reported in Paggio et al. (2018), where we performed head movement detection in a subset of the data only consisting of two speakers. The best result was obtained in that study by a Multilayer Perceptron trained on visual and acoustic features, which achieved 0.75 accuracy and outperformed a classifier trained on monomodal visual features. The performance of the best model in the current study, which applies to the entire dataset, is only about 2% lower, thus showing that our methodology is reasonably robust.

An interesting question is whether approaching the problem in terms of single frames is a good way of approximating what the human annotators did. After all, they were asked to annotate whole head movements, not individual frames.

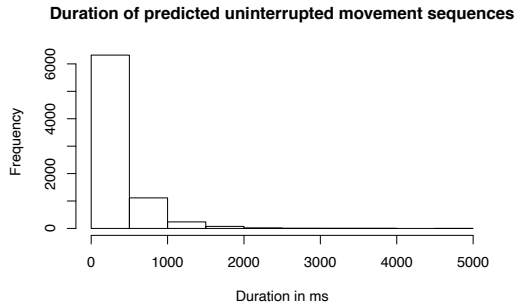

Figure 4: Histogram of the duration of uninterrupted sequences of movement frames predicted by the binary MLP classifier in exp. 4

A way to compare the results of the frame-wise predictions made by the models is to look at the number and duration of uninterrupted movement frame sequences and compare them with the gold standard. The total number of movements predicted by the best binary classifier is 7,782, and their mean duration is 291.25 ms (SD: 360.91). In comparison to the annotated movements, the classifier detects many more but shorter ones. In figure 4 we visualise the whole distribution of the duration of the predicted movements. If we compare it with the histogram in figure 2, we can clearly see that the classifier tends to find many more shorter movements (up to 500 ms), and even though the distribution is also left-skewed, the maximum duration of 4,880 is considerably shorter than the longest movement in the gold standard. There may be several explanations for these differences, e.g. the fact that annotators may have seen a sequence of movements as an uninterrupted repeated gesture of a certain kind rather than separate individual ones.

Looking at the feature combinations used in the experiments, the results confirm the fact that combining the three movement derivatives in the training reliably improves detection and classification for all the models. It can be discussed, however, whether all the values currently used in the vectors are in fact necessary. Having a representation of velocity, acceleration and jerk not only in terms of polar coordinates but also in terms of cartesian coordinates is

redundant since such representations are equivalent. We repeated some of the experiments without the inclusion of polar coordinates. Only the MLP classifier was not adversely influenced by this and became even marginally better. The linear classifiers, on the other hand, performed not any better than the baseline without the polar coordinates.

The role played by the acoustic and word features, on the contrary, is not totally clear in that they only add marginal gains to the F1 scores obtained by the models and in some cases even harm them. It is possible that the speech signal is superfluous, but also that we have not found the most efficient way to combine those features with the visual ones. More research is needed to understand this.

Finally, as we noted the performance of the classifiers varies depending on the speaker. A first analysis of the data indicates that the factors which might influence the results in this direction are the types of head movement performed by the speakers as well as whether the speaker is standing on the same side during the two conversations or not.

7. Conclusions and future work

In conclusion, we have shown that head movements can be detected in unseen speaker data by an MLP classifier trained with multimodal data including movement and acoustic features. The results achieved by this classifier perform at state-of-the-art level. When the same method is applied to the classification of four different types of head movement in the same data, the performance decreases.

In order to develop the present work further, we can investigate different approaches. Firstly, we plan to add more features from OpenPose: the position of ears and chin, for example, might be helpful to add to the position of the nose for some of the head movements. An alternative to OpenPose, or a method that we would like to use in combination with it, could be found in computer vision techniques that identify changing head positions as proposed in Ruiz et al. (2018), who trained a multiloss Convolutional Neural Network on a synthetically created dataset in order to predict yaw, pitch and roll from image intensities.

Secondly, we intend to investigate different ways to use acoustic and word features, either by adding more features or by using them in more selective ways for specific head movement classes.

Thirdly, we would like to analyse the extent to which the depth of the neural network contributes to the results by testing different numbers of layers. Furthermore, we would like to experiment with sequential models such as Recurrent Neural Networks (RNN), which are often used to analyse video sequences and might therefore predict gestures more precisely than the classifiers we have tested until now. In that connection, it would also be interesting to experiment with an architecture in which representations are learnt separately for each feature by different networks and then concatenated into one vector.

Finally, we want to carry out a more precise comparison of the movements predicted and the annotated ones by making the predictions readable by the ANVIL gesture annotation tool.

8. Ethical considerations

We have obtained written permission by the participants to use the videos for research purposes specific to the project within which the recordings were obtained. Therefore, we are making all the features extracted from the corpus available together with the code we have used to train and test the classifiers. However, we do not share the videos or the transcriptions from the corpus because of privacy and data protection issues.

9. References

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In Jean-Claude Martin, et al., 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.

Allwood, J. (1988). The Structure of Dialog. In Martin M. Taylor, et al., editors, *Structure of Multimodal Dialog II*, pages 3–24. John Benjamins, Amsterdam.

Ambrazaitis, G. and House, D. (2017). Acoustic features of multimodal prominences: Do visual beat gestures affect verbal pitch accent realization? In Slim Ouni, et al., editors, *Proceedings of The 14th International Conference on Auditory-Visual Speech Processing (AVSP2017)*. KTH.

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

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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

Frid, J., Ambrazaitis, G., Svensson-Lundmark, M., and House, D. (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)*, number 141, pages 4–9, Copenhagen, September 2016. Linköping University Electronic Press, Linköpings universitet.

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

Germesin, S. and Wilson, T. (2009). Agreement detection in multiparty conversation. In *Proceedings of ICMI-MLMI 2009*, pages 7–14.

Granström, B. and House, D. (2005). Audiovisual representation of prosody in expressive speech communication. *Speech Communication*, 46(3):473–484, July.

Hadar, U., Steiner, T., Grant, E., and Clifford Rose, F. (1983). Head Movement Correlates of Juncture and Stress at Sentence Level. *Language and Speech*, 26(2):117–129, April.

Jongejan, B., Paggio, P., and Navarretta, C. (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*, number 141, pages 10–17. Linköping University Electronic Press, Linköpings universitet.

Jongejan, B. (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. European Language Resources Distribution Agency.

Kapoor, A. and Picard, R. W. (2001). A real-time head nod and shake detector. In *Proceedings of the 2001 Workshop on Perceptive User Interfaces*, PUI '01, pages 1–5, New York, NY, USA. ACM.

Kendon, A. (2004). *Gesture*. Cambridge University Press.

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

Loehr, D. P. (2007). Aspects of rhythm in gesture and speech. *Gesture*, 7(2).

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

Morency, L.-P., Sidner, C., Lee, C., and Darrell, T. (2005). Contextual recognition of head gestures. In *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*.

Morency, L.-P., Quattoni, A., and Darrell, T. (2007). Latent-dynamic discriminative models for continuous gesture recognition. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

Morency, L.-P. (2009). Co-occurrence graphs: contextual representation for head gesture recognition during multiparty interactions. In *Proceedings of the Workshop on Use of Context in Vision Processing*, pages 1–6.

Navarretta, C., Ahlsén, E., Allwood, J., Jokinen, K., and Paggio, P. (2011). Creating Comparable Multimodal Corpora for Nordic Languages. In *Proceedings of the 18th Conference Nordic Conference of Computational Linguistics*, pages 153–160, Riga, Latvia, May 11-13.

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

Paggio, P., Allwood, J., Ahlsén, E., Jokinen, K., and Navarretta, C. (2010). The NOMCO multimodal nordic resource - goals and characteristics. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Paggio, P., Navarretta, C., and Jongejan, B. (2017). Automatic identification of head movements in video-recorded conversations: Can words help? In *Proceedings of the Sixth Workshop on Vision and Language*, pages 40–42, Valencia, Spain, April. Association for Computational Linguistics.

Paggio, P., Jongejan, B., Agirrezabal, M., and Navarretta, C. (2018). Detecting head movements in video-recorded dyadic conversations. In *Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct*, pages 1–6.

Ruiz, N., Chong, E., and Rehg, J. (2018). Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2074–2083.

Ruttkay, Z. and Pelachaud, C. (2006). *From Brows to Trust: Evaluating Embodied Conversational Agents*. Human–Computer Interaction Series. Springer Netherlands.

Swerts, M. and Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, 36(2):219–238.

Tan, W. and Rong, G. (2003). A real-time head nod and shake detector using HMMs. *Expert Systems with Applications*, 25(3):461–466.

Thorsen, N. (1980). Neutral stress, emphatic stress, and sentence Intonation in Advanced Standard Copenhagen Danish. Technical Report 14, University of Copenhagen.

Wei, H., Scanlon, P., Li, Y., Monaghan, D. S., and O'Connor, N. E. (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.

Wu, Y. and Huang, T. S. (1999). Vision-based gesture recognition: A review. In *International Gesture Workshop*, pages 103–115. Springer.

Yngve, V. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society*, pages 567–578.