Multimodal Detection and Classification of Head Movements in Face-to-Face Conversations: Exploring Models, Features and Their Interaction

Manex Agirrezabal[1], *Patrizia Paggio*[1,2], *Costanza Navarretta*[1], *Bart Jongejan*[1]

[1]Centre for Language Technology, University of Copenhagen, Denmark
[2]University of Malta

{manex.aguirrezabal,paggio,costanza,bartj}@hum.ku.dk

Abstract

In this work we perform multimodal detection and classification of head movements from face to face video conversation data. We have experimented with different models and feature sets and provided some insight on the effect of independent features, but also how their interaction can enhance a head movement classifier. Used features include nose, neck and mid hip position coordinates and their derivatives together with acoustic features, namely, intensity and pitch of the speaker on focus. Results show that when input features are sufficiently processed by interacting with each other, a linear classifier can reach a similar performance to a more complex non-linear neural model with several hidden layers. Our best models achieve state-of-the-art performance in the detection task, measured by macro-averaged F1 score.

Index Terms: Head Movement Detection, Multimodal Corpora, Visual and Speech Features

1. Introduction

Head movements are an essential part of face-to-face communication. They have many functions. For example, they allow speakers to effectively give and receive feedback, and thus, they contribute to rapport and mutual comprehension between them. They also support turn exchange as well as speakers' management of their own communicative behaviour, e.g. for lexical search. Therefore, correct detection and interpretation of head movements is crucial for the development of multimodal interfaces that can communicate with users in as natural a way as possible.

The aim of this paper is to experiment with multimodal detection and classification of head movements in a corpus of conversations, and to investigate in particular the importance of different features for these tasks. We detect head movements from raw video data by obtaining key points of relevant body positions as well as pitch and intensity of the relevant speaker. The body positions included are nose, neck and mid hip. We further process this information by calculating first, second and third order derivatives of these positions and audio features, resulting in velocity, acceleration and jerk values. We perform a feature effect analysis and explore whether the combination of features improves previous results. Since our results are obtained on a specific dataset of Danish conversations, they cannot be compared against publicly available benchmarks. Therefore, in addition to providing our own baseline, we also loosely compare the results of our models against state-of-the-art for head movement detection obtained on different data. Besides, we provide some evidence for which features may be the most effective in the task of head movement detection.

The article is structured as follows. We begin by reviewing related work on head movement detection. After that we introduce the corpus used in our experiments. Then, we introduce the tasks and discuss the features, models and validation procedure that constitute our methodology. We continue by presenting and discussing the results of the various models. We conclude the paper with some final remarks and suggestions for possible future directions.

2. Related work

Head movement detection can be performed with relatively high accuracy from data tracked through sensors (Kapoor and Picard, 2001; Tan and Rong, 2003; Wei et al., 2013; Severin, 2021). Doing the same from raw video data is a harder task (Wu and Huang, 1999; Gavrila, 1999), and may require optimal light and background conditions, e.g. when using optical motion flow (Zhao et al., 2012).

Promising results were achieved in detecting head movements in a Swedish corpus of read news (Ambrazaitis and House, 2017; Frid et al., 2017) using velocity and acceleration. The task was formulated in this study in terms of predicting for each word whether or not it was accompanied by a movement of the head. Velocity and acceleration were also used in Jongejan (2012), and enriched with jerk in Jongejan et al. (2017) to detect head gestures in video-recorded free conversations. The detected movements correlated well with the manual annotations at the onset, but generated a high number of false positives.

Work relying on acoustic speech features include Germesin and Wilson (2009), where pitch and energy of voice were combined with word, pause and head pose information to identify agreement and disagreement signals in meeting data, as well as Paggio et al. (2018) and Paggio et al. (2020), where movement features were considered together with pitch, intensity and the presence of silence to identify head movements in conversational data. Evidence for a multimodal approach to the task comes from linguistic and psycho-linguistic research on audio-visual prominence, which has described the close relationship there is between facial beats and acoustic prominence (Granström and House, 2005; Swerts and Krahmer, 2008; Ambrazaitis and House, 2017).

Several studies, in fact, have looked at ways to combine visual and acoustic or language features to detect head movements. For example, in Morency et al. (2005), features from the dialogue context were used together with visual features to predict feedback nods and shakes in human-robot interactions. Speech cues such as the occurrence of specific words and pauses were added in Morency (2009) to improve the detection of head gestures in a vision-based Latent-Dynamic Conditional Random Field (LDCRF) model. LDCRF was found to be the best performing model for head movement recognition in a range of different datasets (obtained from human-robot, human-widget or human-agent interactions) with reported accuracy rates be-

tween 0.75 and 0.8 (Morency et al., 2007), probably due to its ability to deal with the unsegmented nature of movement sequences.

Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) have recently been used to predict head nods and turn taking in the IEMOCAP human-human dyadic conversations (Türker et al., 2018) and to recognise human action from videos (Ullah et al., 2018). In particular for head nod recognition, Türker et al. (2018) report an F1 score of 63.59 obtained by an LSTM-RNN trained on multimodal vectors of non-verbal and acoustic (spectral and prosodic) features. The dataset used seems to contain head nods as the only annotated head movement.

In conclusion, the task of head gesture recognition from raw video data has been studied in specific communicative situations, for limited datasets and with a focus on specific movements such as nods and shakes. More work is needed to validate and further develop state-of-the-art methods on additional datasets and settings in order to validate results obtained from multimodal English data on data from speakers of different languages interacting in different situations.

In this paper, we present a study in which a range of different models are applied to the tasks of detecting and classifying head movement in conversational Danish data and determining the type of movement. As in previous work, we train our models on a combination of visual and acoustic features. Given the lack of benchmarks for the task, it is difficult to compare our results to the state of the art. However, compared to Paggio et al. (2020), the best results we obtain on the same dataset are similar, but achieved by relying on a polynomial kernel rather than a multilayer perceptron to manage complex feature combinations. The results from the binary classification task (detecting head movement) are similar to those reported in Türker et al. (2018) for head nod detection using a different dataset and a more complex classifier.

3. The corpus

The data used in this study come from the NOMCO corpus of annotated first acquaintance conversations in Danish (Paggio et al., 2010; Paggio and Navarretta, 2016). NOMCO consists of twelve dyadic conversations that took place in a recording studio and involved twelve different speakers (six females and six males). Each speaker is audio-recorded while interacting with a person of the same gender and one of the opposite one. The speakers do not know each other, and the purpose of the conversation is to get acquainted. The two speakers are standing in front of each other so that the entire body is visible. The conversations were recorded using three cameras.

Figure 1: *Screenshot from one of the NOMCO conversations: split view*

In this work, we use the recordings from two frontal cameras, which were combined into one video as shown in Fig. 1.

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

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

The duration of each conversation is about five minutes, which results in about one hour of total interaction.

The annotation of the corpus includes, in addition to the speech transcription and many other annotation types, a specification of temporal segments corresponding to different types of head movement that was obtained manually following the MUMIN annotation scheme (Allwood et al., 2007). Head movements were segmented by defining the start and end of each movement and assigning the resulting segment to the right type. If the annotator found a sequence of different movements without a pause in between, say a nod followed by a shake, the two movements were annotated as separate, adjacent segments. If the same movement was repeated without pauses, however, it resulted in one segment, e.g. a repeated nod. The inter-coder agreement reached on identification and classification of head movements was a Cohen's *kappa* score between 0.72 and 0.8 (Navarretta et al., 2011).

Table 1 shows the distribution of head movements both in terms of uninterrupted sequences and single frames, as well as the number of frames not containing any movement. As can be seen, non-movement frames are about twice as many as those that have been annotated as showing movement. The *Other* category corresponds to six different types namely *HeadBackward, HeadForward, SideTurn, Tilt, Waggle* and *HeadOther*. There is of course a fair amount of speaker variation in both number and types of movements produced. As for the duration of the head movements, it is 934.78 ms on average (sd: 579.44). Although most movements are shorter than 1500 ms, there is a long tail of outliers with a maximum duration of up to 7,080 ms, in many cases due to repeated movements.

4. Methodology

We experiment with two tasks: i. detecting presence or absence of head movement (binary classification) and ii. determining head movement type given the four classes *None*, *Nod*, *Shake* and *Other* (multinomial classification). For both tasks, we segment videos into frames and consider the aligned audio signal. Predictions are done for each frame.

Visual features are extracted using the OpenPose library(Cao et al., 2018). We extract positions for the nose, the neck and the mid hip. For each of these positions we include the Cartesian (x and y) coordinates and a weight feature calculated by OpenPose. 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. For each derivative, we use the x and y coordinates as well as polar (radius *r* and angle *clock*) coordinates, where the radius of the position is the same as its Euclidean norm $\sqrt{x^2 + y^2}$. This gives 15 features per body part, for a total of 45 visual features.

Acoustic features were obtained by extracting pitch and

intensity values using the Praat tool for phonetic analysis (Boersma and Weenink, 2009). In both cases, the extraction was done with a time step of 0.04s, which outputs 25 measurements per second. For pitch extraction, the same range of 75-600 was used for both genders. While using this range means allowing for a considerable amount of noise, experiments done with different ranges resulted in too many missing values. In general, it must be noted that the quality of the acoustic signal is far from optimal since the recordings were obtained with external microphones hanging from the ceiling. For practical reasons, pitch and intensity were treated as if they were in the Cartesian coordinate system, and therefore, the derivatives include both Cartesian coordinates (x and y) but also polar coordinates (r and $clock$). Similarly to what was done for the raw visual data, acoustic measurements were used to compute three derivatives, i.e. velocity acceleration and jerk, corresponding to four different features for each derivative, which, added to the two primary ones, leaves us with 14 acoustic features.

To test the way acoustic derivatives may be used to model our data, a preliminary statistical analysis was conducted using the R software (R Core Team, 2020). Data from all speakers were pooled together, and generalised linear models were created to predict the probability of a head movement in a video frame given several movement and sound derivatives, first separately, and then in combination. The 'glm' function[1] was used to fit a number of binomial models to predict the probability of head movement, which was expressed as a binary value (1 for movement and 0 for non-movement). The best performing model found small but significant effects for velocity of pitch and acceleration of intensity in combination with nose tip acceleration coordinates, thus providing some support for the use of the acoustic derivatives.

To model the co-occurrence between head movements and verbalisation in the data, we also added a feature that encodes for each frame whether the speaker in focus is actually speaking or not.

In sum, we obtained a representation for each frame that contains 45 visual features, 14 acoustic ones and one regarding the presence or absence of verbalisation. The resulting data were used to train our models.

As mentioned above, our corpus contains videos with 12 speakers. We trained 12 speaker-independent models and evaluated them using a leave-one-out cross-evaluation as follows: For each speaker being tested, data from the two videos in which the speaker is recorded are kept for testing and data from the remaining eleven speakers are used for training. Test results are given in terms of macro-averaged F1 values across the 12 runs.

The following classifier types were used to train models using various feature combinations: i. Logistic Regression (LR), which is an example of a simple model, ii. Linear Support Vector Machine (LINEARSVC), which was used by several earlier studies for head movement detection, iii. Multilayer Perceptron (MLP) with four layers, as an example of a non-linear classifier, iv. Conditional Random Field (CRF) and v. Latent-Dynamic Conditional Random Field (LDCRF), for both of which we reused code made available by the authors of Morency (2009)[2].

In addition to reporting F1 values obtained by always choosing the most frequent label (MF), we also consider a baseline in which only velocity derivatives for the nose, neck and mid hip are used by the classifiers. Then we conduct three experiments using features in different ways.

In Exp 1 all the visual and sound features are used. Here we expand on the array of visual features used in previous work, e.g. Paggio et al. (2020), by adding some based on neck and mid hip positions. We expect the new features to help in distinguishing between movements of different kind and in isolating the head from the rest of the body.

An analysis of the effects provided by all the features as used by the LR model showed that many of them seemed to contribute very little to the analysis (Fig. 2). Therefore, in Exp 2 we experiment with removing all the features with an effect lower than 1. That leaves a total of 17 features, notably including some position and derivative values from all three body parts, as well as pitch velocity and jerk, but no intensity feature.

The feature analysis also shows that the radius of the nose acceleration and jerk are the two most predictive features, with an effect that is substantially higher than any of the others. To model how these two features interact with the others, in Exp 3 we train the models using a polynomial kernel and we reduce the feature space to the initial one by using Principal Component Analysis (PCA). Only LR, LSVM and MLP classifiers were tested with this configuration.

5. Experiment results and discussion

The results of the experiments are shown in Table 2 for binary and Table 3 for multinomial classification. In general, all the models do much better in the binary classification task. For both tasks, the results of Exp 1 are better than those obtained with the BS. We also see that the results of Exp 2 are relatively low, which seems to indicate that even features that contribute little individual gain can be used in interaction with the others.

In the first task, the best F1 score is obtained by the LR in combination with the polynomial kernel. The MLP classifier yields practically the same score using all the features. In the second task, the best F1 scores are again obtained using the kernel, this time by the LSVM classifier. The MLP trained with all the features yields slightly lower results.

Surprisingly, the CRF and LFCRF models do not perform well in either task, possibly because the dataset is too small for them to work effectively.

If we compare the performance of our models with the state of the art Türker et al. (2018), our models perform similarly in the binary classification task. However, our task is not exactly the same since Türker et al. (2018) detect head nods in a dataset in which those are the only head movement type annotated, whereas we are trying to detect different types of head movement. Our task is, in other words, harder. This is reflected in the much lower results obtained in the multinomial classification experiments. A preliminary analysis showed, in particular, that shakes are much harder to classify than nods. More research is needed to understand why and to improve the results.

To position our work with respect to the best results obtained on more constrained dialogue types, we also trained CRF and LDCRF models from Morency et al. (2007) with our own data. The results are significantly lower than reported in the original paper. It might be that the model developed by Morency et al. (2007) is better suited for lab settings in which the conversations are rather constrained (human-robot, human-widget or human-agent), but further analysis is required to corroborate this.

When we look at the predicted movements as entire movement sequences rather than independent single frames, we see

[1] https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm

[2] https://sourceforge.net/projects/hcrf/

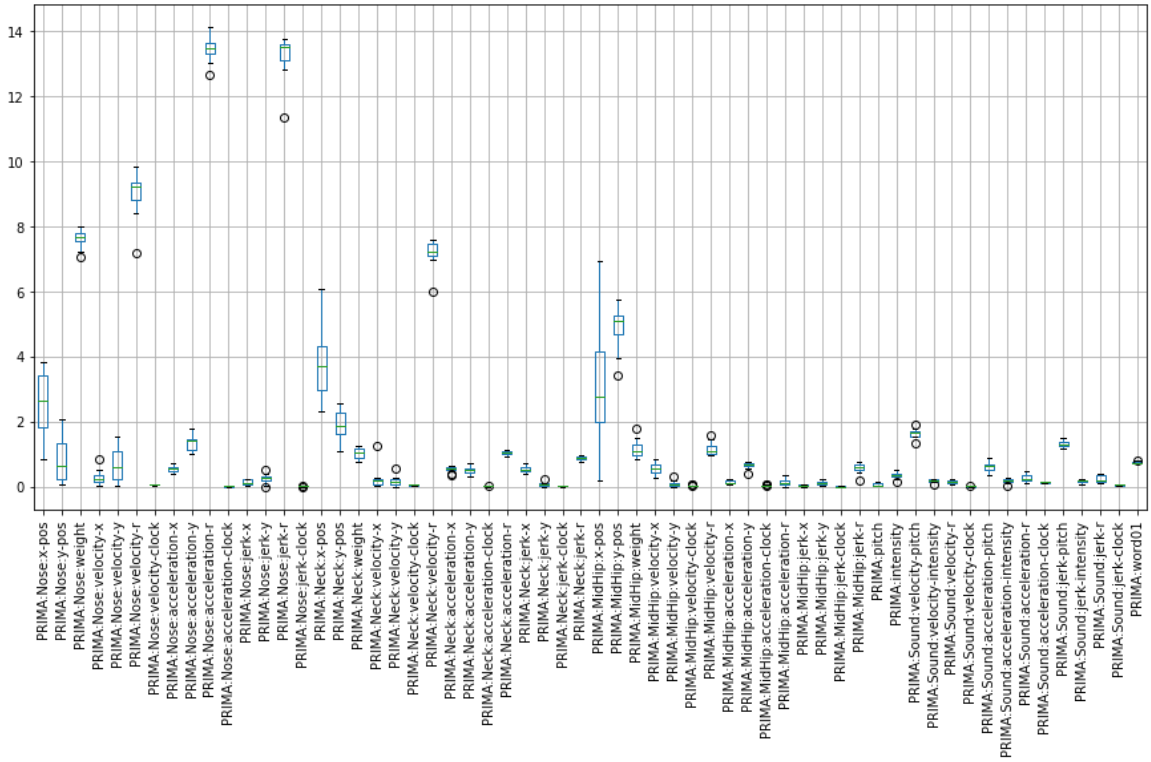

Figure 2: *Visualisation of the effect of each feature (absolute values) in predicting presence or absence of head movements.*

Table 2: *Macro F1-score values for binary classification*

		MF	LR	LSVM	MLP	CRF	LDCRF
BS	Nose, neck and mid hip velocity	0.3871	0.6033	0.5923	0.6595	0.4474	0.5500
Exp 1	All features	0.3871	0.6807	0.6725	0.6842	0.4252	0.3836
Exp 2	Features w/ effect>1	0.3871	0.6699	0.6580	0.6567	0.3441	0.3178
Exp 3	Polynomial kernel	0.3871	**0.6848**	0.6782	0.6450	-	-

that 86% of them were predicted with various degrees of overlap, and that movement on-sets are more easily detected than offsets. A high number of non-existing movements are also, however, incorrectly predicted, probably due to the fact that only movements that have a function in the conversations were coded by the human annotators.

Now considering the features used to train the models, we see that the full set of features, combined using either a polynomial kernel or hidden layers of a MLP, enhances the performance of both binary and multinomial classifiers.

With regards to the relative importance of features in predicting the presence or absence of movement, we can see in Fig. 2 that when considering the key points from the nose, the x and y values do not contribute much no matter which derivative is used. LR requires x and y to be combined, and that is why the *r* (radius) feature, which encodes the Euclidean norm of a vector (x, y), has a considerable importance if compared to other features. It seems that the model needs to know how large the vector is to assess whether there is a movement or not. In relation to this, we can observe that the angle of the (x, y) vector does not contribute to a better performance (*clock* feature). If we consider the features from neck positions, our analysis shows that the Euclidean norm of the velocity is important. The positions of the neck and mid hip can also be seen as modestly important. We hypothesise that the Euclidean norm might be a valuable feature because it makes the values of x and y positive and combines them.

The second order polynomial kernel shows promising results in some classifiers and not so promising ones for some others. On the one hand, both linear classifiers (LR and LSVM) reach a similar performance to the MLP with all the features, but on the other hand, the performance of the MLP decreases. This suggests that if we process the data sufficiently, then a simple linear classifier can manage the task. This is interesting if we consider that the LR takes less than 10 minutes to train in all our experiments, even when features are combined with a polynomial kernel. Its efficiency makes it a relevant competitor to more complex classifiers, such as the MLP, which needs 45 minutes to train. The LSVM, in turn, takes around 15 minutes to train in the binary task and around 45 minutes in the multiclass task.

We also analysed feature importance for different types of head movements. There is a general trend that in both general feature effects and also in the effect for nod or jerk movements, the y position is more important than the x position for both the nose and the neck. If we consider head shakes, the x position seems to be more relevant than y, with a relatively large margin. Considering that x and y positions are the ones that encode

Table 3: *Macro F1-score values for multinomial classification*

		MF	LR	LSVM	MLP	CRF	LDCRF
BS	Nose, neck and mid hip velocity	0.1936	0.2726	0.2675	0.3301	0.2811	0.2870
Exp 1	All features	0.1936	0.3330	0.3174	0.3766	0.2341	0.2304
Exp 2	Features w/ effect>1	0.1936	0.3023	0.2924	0.3365	0.1549	0.1721
Exp 3	Polynomial kernel	0.1936	0.3881	**0.3893**	0.2899	-	-

the horizontal and vertical axes, respectively, this makes sense, especially because a head shake involves a slight rotation of the head in the horizontal axis.

6. Conclusion and future research

In this work we experimented with a number of features and combinations for the detection and classification of head movements in conversations. The performance of the models is at state-of-art level. It is achieved, however, on a relatively small dataset annotated with many different head movement types. We tested several features and models and observed that when linear models are given sufficiently processed information, they reach similar performance to more complex ones. We also conducted a feature analysis that showed the importance of the more complex features combining information from the more primary ones, but also that the complete feature set yielded the best results in combination with the simplest features. We made our code available on Github.[3]

This work could be further developed in several ways. Regarding models, we have used rather simple Machine Learning models and in the future we would like to explore further. Based on similar research, we expect that more complex neural networks, such as Long Short-Term Memory or Convolutional Neural Network models, might improve the current performance. For example, Slowfast CNNs (Zhang et al., 2021) have shown good performance for isolated gesture recognition, but also for head movement detection (Xie et al., 2021).

In our feature analysis acoustic features did not perform as well as we were expecting. We experimented with speaker-level normalisation of the acoustic measures to investigate whether individual variation had a negative effect on the results, but normalisation did not improve the results. Therefore, we suspect that the poor quality of the acoustic signal might be the reason why acoustic features do not seem to improve the classification results.

It would be very interesting to incorporate data from other similar devices or sensors, such as Human Activity Recognition data, and then, use a Transfer Learning approach to assist our current models. This approach has been successfully used in different works, for instance, (Gashi et al., 2021) where publicly available Human Activity Recognition data was used to enhance a head gesture recognition model. In recent work, transfer learning was also used to combine information from different parts of the body (Zhong et al., 2022), and we believe that a similar approach could be used to combine the keypoints from OpenPose that we currently use.

Last but not least, we are currently using the best of our models to provide a rough head movement annotation in a newly created corpus of online zoom meetings in English. The annotation process will provide an excellent test-bed for an evaluation of how useful the models' output is in a realistic corpus annotation scenario. The final annotated data will be made available to the research community.

[3] https://github.com/kuhumcst/head_movement_detection

7. Acknowledgements

We would like to acknowledge the support of the international research network GEstures and Head Movements in language (GEHM), which is funded by the Danish Research Council (grant number: 9055-00004B).

8. 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. *Multimodal Corpora for Modelling Human Multimodal Behaviour*, 41(3-4):273–287.

Ambrazaitis, G. and House, D. (2017). Acoustic features of multimodal prominences: Do visual beat gestures affect verbal pitch accent realization? In Ouni, S., Davis, C., Jesse, A., and Beskow, J., editors, *Proceedings of The 14th International Conference on Auditory-Visual Speech Processing (AVSP2017)*, pages 89–94, Stockholm, Sweden. 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*.

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. Linköping University Electronic Press, Linköpings universitet.

Gashi, S., Saeed, A., Vicini, A., Di Lascio, E., and Santini, S. (2021). Hierarchical classification and transfer learning to recognize head gestures and facial expressions using earbuds. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, page 168–176, New York, NY, USA. Association for Computing Machinery.

Gavrila, D. M. (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.

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.

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.

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.

Morency, L.-P. (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.

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., Sidner, C., Lee, C., and Darrell, T. (2005). Contextual recognition of head gestures. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 18–24.

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.

Paggio, P., Agirrezabal, M., Jongejan, B., and Navarretta, C. (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).

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., 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.

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

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Severin, I.-C. (2021). Head gesture-based on imu sensors: a performance comparison between the unimodal and multimodal approach. In *2021 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4.

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.

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

Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., and Baik, S. W. (2018). Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE Access*, 6:1155–1166.

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.

Xie, J., Zhang, B., Chepinskiy, S. A., and Zhilenkov, A. A. (2021). A dynamic head gesture recognition method for real-time human-computer interaction. In Liu, X.-J., Nie, Z., Yu, J., Xie, F., and Song, R., editors, *Intelligent Robotics and Applications*, pages 235–245, Cham. Springer International Publishing.

Zhang, X., Tie, Y., and Qi, L. (2021). Slowfast convolution lstm networks for dynamic gesture recognition. In *2021 3rd Asia Pacific Information Technology Conference*, APIT 2021, page 59–63, New York, NY, USA. Association for Computing Machinery.

Zhao, Z., Wang, Y., and Fu, S. (2012). Head movement recognition based on Lucas-Kanade algorithm. In *2012 International Conference on Computer Science and Service System*, pages 2303–2306. IEEE.

Zhong, J., Li, J., Lotfi, A., Liang, P., and Yang, C. (2022). An incremental cross-modal transfer learning method for gesture interaction. *Robotics and Autonomous Systems*, 155:104181.