Towards classification of head movements in audiovisual recordings of read news

Johan Frid
Lund University
Humanities Laboratory,
Lund University, Sweden
johan.frid@humlab.lu.se

Gilbert Ambrazaitis
Linguistics and Phonetics,
Centre for Languages and Literature,
Lund University, Sweden
gilbert.ambrazaitis@ling.lu.se

Malin Svensson-Lundmark
Linguistics and Phonetics,
Centre for Languages and Literature,
Lund University, Sweden
malin.svensson_lundmark@ling.lu.se

David House
Department of Speech, Music and Hearing, KTH, Sweden
davidh@speech.kth.se

Abstract

In this paper we develop a system for detection of word-related head movements in audiovisual recordings of read news. Our materials consist of Swedish television news broadcasts and comprise audiovisual recordings of five news readers (two female, three male). The corpus was manually labelled for head movement, applying a simplistic annotation scheme consisting of a binary decision about absence/presence of a movement in relation to a word. We use OpenCV for frontal face detection and based on this we calculate velocity and acceleration features. Then we train a machine learning system to predict absence or presence of head movement and achieve an accuracy of 0.892, which is better than the baseline. The system may thus be helpful for head movement labelling.

1 Introduction

This study was conducted in the context of a research project on multimodal, or audiovisual, prosodic prominences and their utilization for information structure coding. In particular, the project investigates how head and eyebrow movements interact with sentence-level pitch accents (also referred to as focal accents in the Swedish prosody research community, cf. Bruce (1977), Bruce & Granström (1993)). A crucial part of the project's aim is to explore possible co-occurrences of the three prominence cues (focal accents, head beats and eyebrow beats). For this purpose, annotations of focal accents, as well as head and eyebrow beats are required. Focal accents are assigned to words: functionally, a focal accent lends prominence to a word, and a word can normally only receive one focal accent. Therefore, the domain of interest in the present context is the word, and for that reason, we have decided to define the word as a domain also for annotations of head and eyebrow movements.

One challenge of such a project lies in the annotation of head and eyebrow movements based on video data, which is commonly achieved by means of manual labelling by human annotators. In order to enable future large-scale investigations of multimodal prominence, we are developing automatic methods for the annotation of movements, in this study strictly focusing on head beats.

To this end, we developed a system for training a classifier to recognise head movements in video data. The purpose of the present study is twofold: 1) to see how well we can classify head movements with an automatic classifier, and 2) to identify labelling-related problems. As motivated above, in the present study we use the domain of the word (rather than, e.g., syllable) because of the relation to information structure.

2 Method

Here we describe our procedures for data collection, annotation, video analysis, feature extraction and machine learning.

2.1 Material

Our materials consist of Swedish television news broadcasts and comprise audiovisual recordings of five news readers (two female, three male) from 144 different sessions. The total duration of the recordings is just over 27 minutes and there are about 4200 spoken words in total. There is always only one person present in the video frame at a given time and he/she almost always faces the camera. Hence, face detection is rather straightforward in this material. The frame rate was 25 fps.

2.2 Annotation

This corpus was manually labelled, applying a simplistic annotation scheme consisting of a binary decision about absence/presence of a movement in relation to a word: to this end, the audio-visual data was first segmented at the word level based on the audio data. Praat (Boersma & Weenink, 2014) was used for this purpose. Each segment was also labelled with the actual word spoken.In total, there were 4208 words. There were also 234 sentence- or phrase-internal pauses. These were also annotated and included in the material as they also may be associated with head movements. In total there were 4442 word units. In the rest of this article, we shall refer to them simply as 'words'.

In the next step, ELAN (Wittenburg et al., 2006) was used to determine for each word if there was head movement or not, where ‘presence’ was defined as an event in which the head rapidly changed its position, roughly within the temporal domain of the word. This was done based on the complete audio-visual display.

Our simple annotation scheme (i.e. assigning annotation to words directly) introduces a problem which results in slight discrepancies: As movements may be realized near the border between two adjacent words, or even span two words, the decision as to which of the words should be annotated for the movement in question is not always obvious.

The material was annotated in three different sets by five annotators. Set 1 consisted of 77 sessions (2554 words) and was annotated by annotator 1, Set 2 consisted of 36 sessions (851 words) and was annotated by annotator 2, and finally, Set 3 had 31 sessions (1037 words) and was annotated (independently) by annotators 3-5. For Set 3, an annotation was counted as such in the event of an agreement between at least two annotators ('majority vote'). Furthermore, for Set 3, the absolute agreement (when all three annotators agreed) was 82.7% and Fleiss' κ (Fleiss, 1971) was 0.69.

As it is possible that annotators behave differently, we will look at each annotator group separately as well as the combination of all three sets. For a more detailed discussion of our definition of beat head movements and our other multi-modal annotations (eyebrow beats and verbal prosodic prominence), see Ambrazaitis et al. (2015).

2.3 Video and head movement analysis

For the video analysis we used the frontal face detection functions in the OpenCV library (Viola & Jones, 2001) to detect areas with faces. This method is similar to Zhang et al. (2007). Each frame in the visual speech corpus is analysed, and this gives us an estimate of the location of the face - and head; they are almost equivalent in this context - as coordinates in the x-y plane, as illustrated in Figure 1.

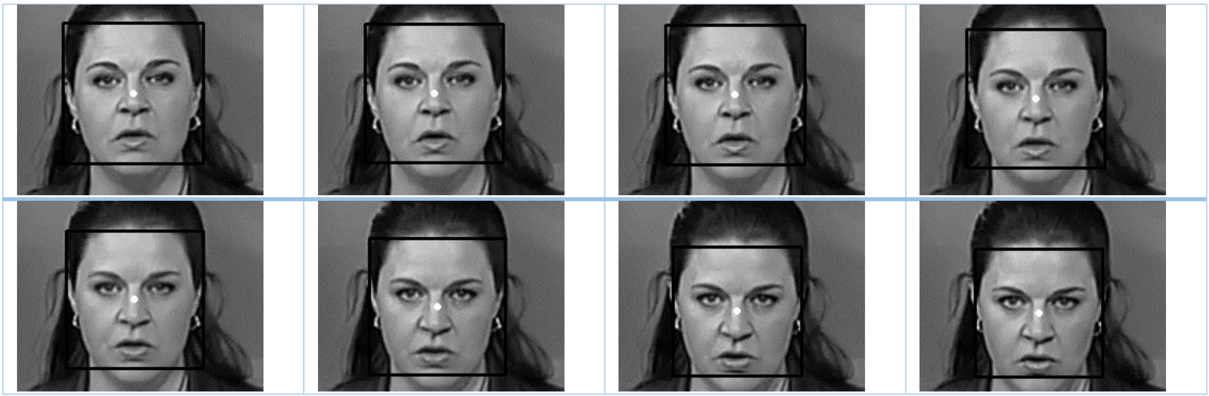

Figure 1. Faces detected in successive frames during a head movement. The black square is the detected face, the white dot (at the center of the square) is the x-y coordinate we use.

The next step is to smooth and calculate velocity and acceleration profiles from the head coordinates. Here we use a method described by Nyström and Holmqvist (2010). We use the Savitzky–Golay (SG) FIR smoothing filter, which makes no strong assumption on the overall shape of the velocity curve and is reported to have a good performance in terms of temporal and spatial information about local maxima and minima (Savitzky & Golay, 1964). Given raw head coordinates this outputs smoothed velocity and acceleration for the x- and y-dimensions separately. Then the total magnitudes of velocity and acceleration are calculated as the Euclidean distance of the x- and y-components. This is shown in Figures 2 and 3, where we also show how we can compare the movement functions with the intervals of our word-related head movement labelling.

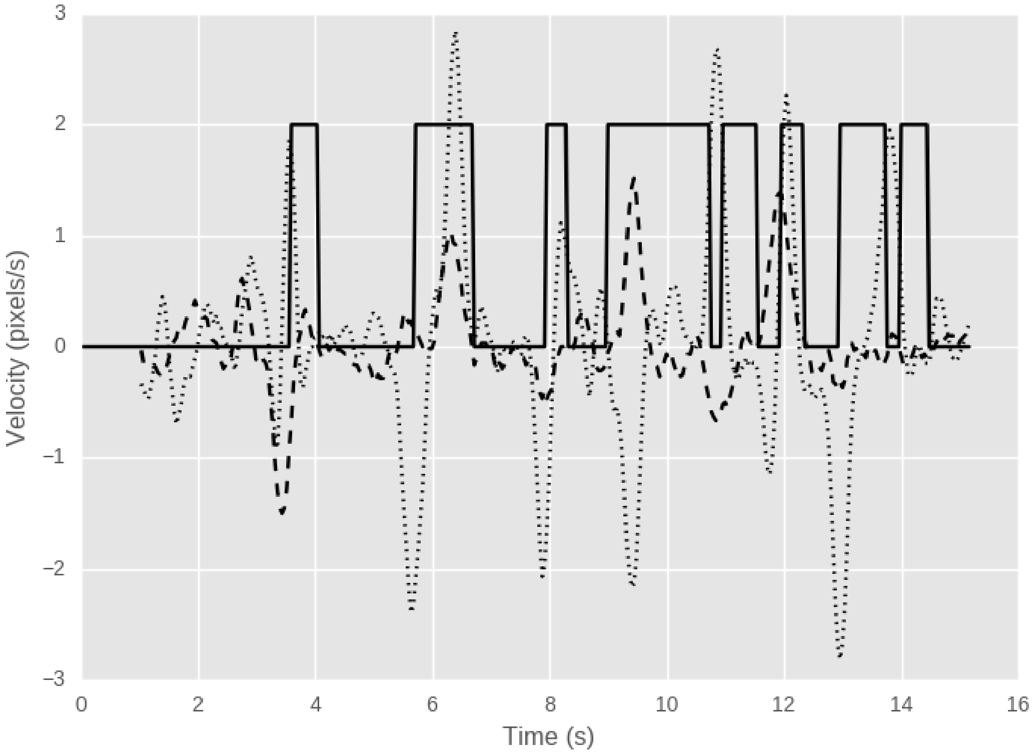

Figure 2. X-velocity (dashed), y-velocity (dotted) and word intervals (solid) as a function of time. The word interval functions have the value 2 in an interval labelled as having movement, and 0 elsewhere.

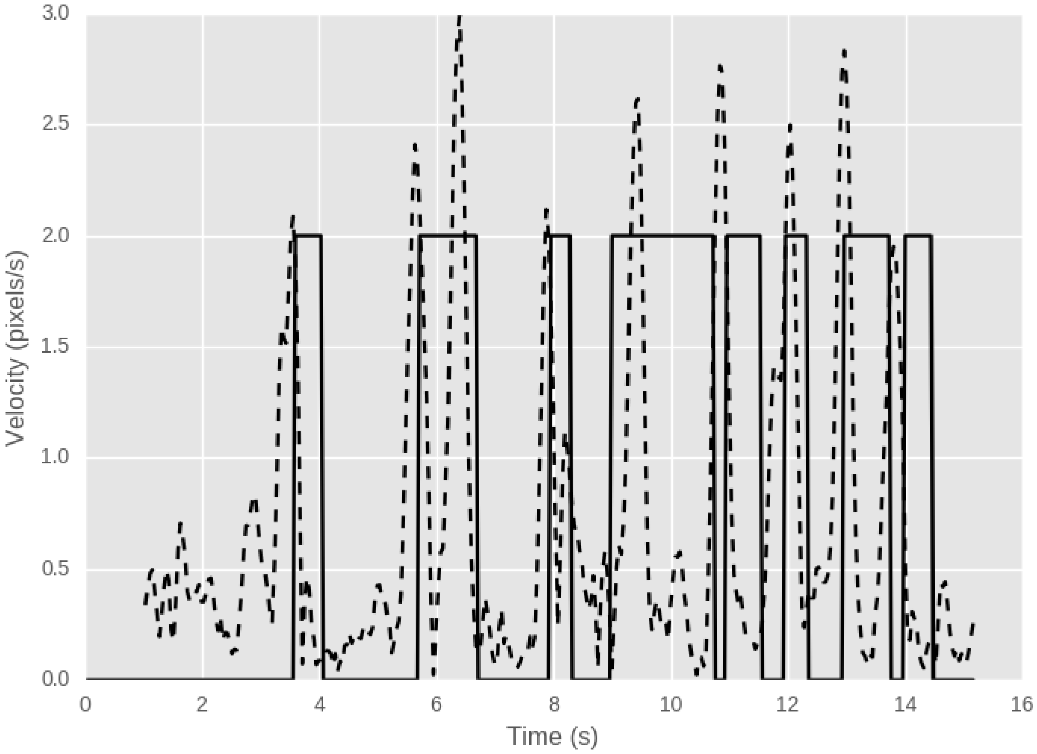

Figure 3.Magnitude of velocity (dashed) and word intervals (solid) as a function of time. The word interval functions have the value 2 in an interval labelled as having movement, and 0 elsewhere.

2.4 Feature extraction

From each of the six curves (x-velocity, y-velocity, x-acceleration, y-acceleration, magnitude of velocity and magnitude of acceleration) we calculate four features per word: average, max, min and amplitude (max-min). This gives us a total of 24 features.

2.5 Classifier

We then trained a classifier by feeding the features into a machine learning algorithm and training it to predict the outcome movement or no movement. We used Xgboost (Chen & Guerstin 2016), which is a popular method in the machine learning community. It is an ensemble of decision trees, where special care is taken to avoid overfitting. The manually annotated data was used to train and evaluate the classifiers.

3 Results

We ran different experiments, both on all data combined and on subgroups, where the data is split on annotator group or news reader. For the annotator groups, we did not perform any experiments on the individual annotations by annotator 3-5; these are collapsed into 'Annotator 3' in the experiments. Finally, as noted in section 2.2, we know that movements can cross word boundaries, we look at a case where we set the label of neighbours of words annotated with 'movement' to 'movement' (regardless of what they were before). In other words, we let the positively annotated words 'leak' into its neighbours. In this way, we may to some extent capture cases where the movements are crossing word boundaries, but where only one of the words has been labelled 'movement'.

All our experiments are run with xgboost, using 10-fold cross-validation. As evaluation measurements we use accuracy (ACC-XGB), F1 score (F1) and area under ROC curve (AUROC). For comparison, we also calculate the accuracy of a baseline classifier (ACC-BL) that always predicts the majority case. We also include the total number of words (N) and the distribution between movement (N(M)) and no-movement (N(NM)) classes in the tables that follow.

3.1 Combined

Results for all words are presented in Table 1. We note that the xgboost classifier outperforms the baseline classifier.

	N	N(M)	N(NM)	ACC -BL	ACC-XGB	F1	AUROC
Combined	4442	720	3722	0.838	0.892	0.624	0.756

Table 1. Results for all words.

3.2 Subgroups: annotators

Results per annotator are in Table 2. We again note that the machine learning classifier is better than the baseline, and also that the evaluation measurements are both lower than and higher than the 'Combined' case (Table 1).

	N	N(M)	N(NM)	ACC -BL	ACC-XGB	F1	AUROC
Annotator 1	2554	395	2159	0.846	0.885	0.556	0.717
Annotator 2	851	96	755	0.887	0.936	0.666	0.803
Annotator 3	1037	229	808	0.779	0.865	0.667	0.786

Table 2. Results per annotator.

3.3 Subgroups: news readers

We show the results split per news reader in Table 3. The xgboost classifier again gets higher accuracy score than the baseline, and compared with the 'Combined' case (Table 1) we again see the results going in both directions.

	N	N(M)	N(NM)	ACC -BL	ACC-XGB	F1	AUROC
Newsreader 1	904	80	824	0.916	0.932	0.542	0.721
Newsreader 2	981	216	765	0.780	0.857	0.672	0.746
Newsreader 3	508	65	443	0.872	0.920	0.661	0.800
Newsreader 4	1318	238	1080	0.819	0.878	0.618	0.755
Newsreader 5	731	121	610	0.834	0.892	0.647	0.785

Table 3. Results per newsreader.

3.4 Neighbours

Finally, the results for the 'Neighbours' case are shown in Table 4. The xgboost classifier again outperforms the baseline if we compare their accuracy. We also note that the difference is much larger than in the 'Combined' case (Table 1).

	N	N(M)	N(NM)	ACC -BL	ACC-XGB	F1	AUROC
Neighbours	4442	1817	2625	0.59	0.738	0.653	0.718

Table 4. Results for all words, with neighbours changed.

4 Discussion

Overall, our system performs better than the baseline, which we take as an indication that it might be useful for labelling new, unknown data.

As regards the differences between the annotators, if the performance of the system had been better for each individual annotator, this would mean that there would be an annotator-dependent pattern that would disfavour grouping all data together. Since this is not the case, we can use different labellers (or labeller groups).

Similarly, if all the results for individual news readers had been better than the 'Combined' case, then our data would not have any generative power. Since this is not the case, we think our method performs well for the general case where all news readers are combined and would be applicable to other news readers.

The classifier may be helpful for head movement labelling in its own right. Moreover, as mentioned in section 2.2 and shown in Figures 2 and 3, our labelling poses some problems for the classifier: we see that there are cases where the peak of the velocity curve crosses the word label function. This means that the head movement occurs right on a word boundary. This is a problem as one word then has been labelled as 'movement' and the other as 'no movement', but both may have large velocity/acceleration. Our 'Neighbours' condition is one attempt to deal with that, and we think that the fact that the improvement over the baseline is larger in this condition indicates that it is useful to look at possibilites beyond the word. We intend to pursue other strategies for this in the future.

Another, less problematic, case is that more than one head movement can co-occur with the same word. Our feature extraction deals with that as it is not dependant upon the number of peaks within a word, just the max, the average etc.

5 Conclusion

We have developed a system for the detection of word-related head movements in audiovisual recordings of read news. The task seems feasible; our data seems to have predictive power. The results show no effects from using individual vs groups of labellers. Furthermore, they show that it is possible to generalize over several different news readers. Labelling at word boundaries causes some issues when head movements occur across boundaries.

Acknowledgements

This work was supported by an infrastructure grant from the Swedish Research Council (Swe-Clarin, 2014–2018; grant number 821-2013-2003), a grant from the Marcus and Amalia Wallenberg Foundation (grant number 2012.0103) and also partially funded by the Bank of Sweden Tercentenary Foundation (grant number P12-0634:1). We also thank our two additional annotators Anneliese Kelterer and Otto Ewald.

Reference

Ambrazaitis, G., Svensson Lundmark, M. & House, D. (2015). Multimodal levels of prominence : a preliminary analysis of head and eyebrow movements in Swedish news broadcasts. In Svensson Lundmark, M., Ambrazaitis, G. & van de Weijer, J. (Eds.) Working Papers in General Linguistics and Phonetics (Proceedings from Fonetik 2015) (pp. 11-16), 55. Centre for Languages and Literature, Lund University.

Boersma, P., Weenink, D. 2014. Praat: doing phonetics by computer [Computer program]. http://www.praat.org/

Bruce, G. 1977. Swedish Word Accents in Sentence Perspective. Travaux de l'institut de linguistique de Lund 12. Malmö: Gleerup.

Bruce, G., B. Granström (1993). Prosodic modelling in Swedish speech synthesis. Speech Communication 13, 63–73.

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining.

Fleiss, J. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), 378-382.

Nystrom, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. Behavior Research Methods, 42, 188-204. doi:10.3758/BRM.42.1.188

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36, 1627-1639.

Viola, P., & Jones, M. J. (2001) Rapid Object Detection using a Boosted Cascade of Simple Features, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. Volume: 1, pp.511–518.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. 2006. ELAN: a professional framework for multimodality research. Proc. of LREC 2006, Fifth International Conference on Language Resources and Evaluation. See also: http://tla.mpi.nl/tools/tla-tools/elan/

Zhang, S., Wu, Z., Meng, H., Cai, L. (2007) Head Movement Synthesis based on Semantic and Prosodic Features for a Chinese Expressive Avatar In: ICASSP 2007, Vol. 4, pp.837-840, 2007.4