Towards classification of head movements in audiovisual recordings of read news

Frid, Johan; Ambrazaitis, Gilbert; Svensson Lundmark, Malin; House, David

2016

Document Version:
Publisher's PDF, also known as Version of record

TOWARDS CLASSIFICATION OF HEAD MOVEMENTS IN AUDIOVISUAL RECORDINGS OF READ NEWS

Johan Frid[1,2], Gilbert Ambrazaitis[2], Malin Svensson Lundmark[2], David House[3]
[1]Humanities Laboratory, Lund University, Sweden
[2]Linguistics and Phonetics, Centre for Languages and Literature, Sweden
[3]Department of Speech, Music and Hearing, KTH, Sweden
{johan,frid|gilbert.ambrazaitis|malin.svensson_lundmark}@ling.lu.se, davidh@speech.kth.se

This study is part of a project investigating levels of multimodal prosodic prominence, as resulting from an interplay of verbal prosody (pitch accents) and visual prosody (head and eyebrow beats). 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, and 2) to identify labelling-related problems and see if it might be useful for the improvement of the labelling process to have access to movement data.

Our materials consist of Swedish television news broadcasts and comprise speech from four new readers (two female) and about 1000 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.

This corpus was previously 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. Then, ELAN was used to determine for each word if there was head movement or not, where 'presence' was defined as the event that the head rapidly changed its position, roughly within the temporal domain of the word. This was done based on the complete audio-visual display, by three annotators independently of each other. Finally, the three annotations were compared, and for the analyses (as well as for the present study), an annotation was counted as such in the event of an agreement between at least two annotators. These annotations constitute the point of departure for the present study. For a more detailed discussion of our definition of beat head movements and our other multi-modal annotations (prosodic prominence), see Ambrazaitis et al (2015).

For the video analysis we used the frontal face detection functions in the OpenCV library 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 equvalent 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 angular velocity and acceleration are calculated as the Euclidean distance of the x- and y-components. This is shown in Figure 2, where we also show how we can compare the movement functions with the intervals of our word-related head movement labelling.

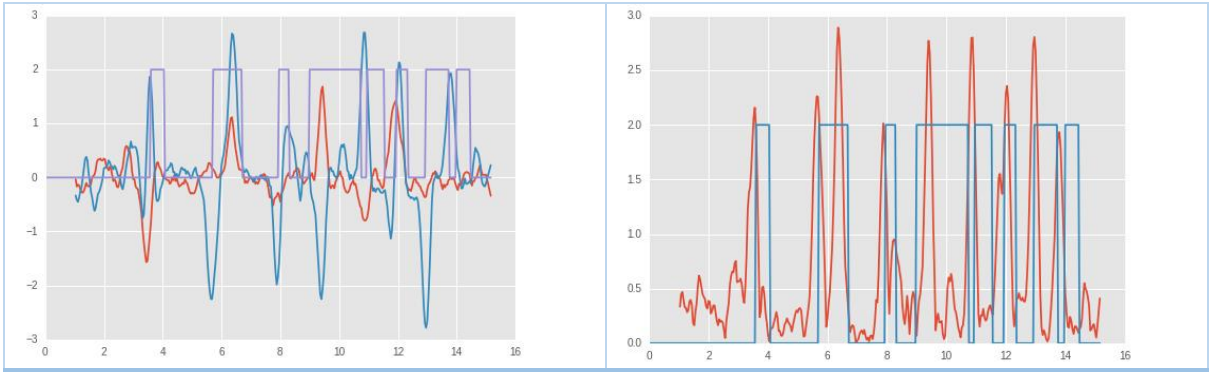

Figure 2. Left: x-velocity (red), y-velocity (blue) and word intervals (purple) as a function of time. Right: angular velocity (red) and word intervals (blue) as a function of time.The word interval functions have the value 2 in an interval labelled as having movement, and 0 elsewhere.

From each of the six curves (x-velocity, y-velocity, x-acceleration, y-acceleration, angular velocity and angular acceleration) we calculate four features per word: average, max, min and amplitude (max-min). We then trained a classifier by feeding the features into a machine learning algorithm. Our test corpus contains 1047 words. Of these, 818 words are labelled as not having head movement (about 78%), and 229 are labelled as having head movement. We ran a 10-fold cross-validation using xgboost (Chen & Guerstin 2016) and this gave us 85% correctly classified words. We thus perform better than the 'majority' vote, which would have assigned 'no movement' to all words.

The classifier may be helpful for head movement labelling in its own right. Moreover, as may be evident from Figure 2, 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. By visualising the head movements in this way we might indicate that some labels needs to be adjusted. Our goal for future work is to improve the classifier and to integrate this in a tool which could facilitate head movement labelling.

Acknowledgements
This work was supported by an infrastructure grant from the Swedish Research Council (SWE-CLARIN, 2014–2018; grant number 821-2013-2003) and a grant from the Marcus and Amalia Wallenberg Foundation (grant number 2012.0103).

References

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.

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

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.

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