Investigating visual prosody using articulography

Johan Frid[1][0000-0000-4704-5559], Malin Svensson Lundmark[2],
Gilbert Ambrazaitis[3][0000-0001-5324-3071], Susanne Schötz[4][0000-0003-3898-7835] and
David House[5][0000-0002-4628-3769]

[1] Lund University Humanities Lab, Lund University
[2] Centre for Languages and Literature, Lund University
[3] Department of Swedish, Linnæus University
[4] Logopedics, Phoniatrics and Audiology, Clinical Sciences, Lund University
[5] Department of Speech, Music and Hearing, KTH
✉johan.frid@humlab.lu.se, davidh@speech.kth.se

Abstract. In this paper we describe present work on multimodal prosody by means of simultaneous recordings of articulation and head movements. Earlier work has explored patterning, usage and machine-learning based detection of focal pitch accents, head beats and eyebrow beats through audiovisual recordings. Kinematic data obtained through articulography allows for more comparable and accurate measurements, as well as three-dimensional data. Therefore, our current approach involves examining speech and body movements concurrently, using electromagnetic articulography (EMA). We have recorded large amounts of this kind of data previously, but for other purposes. In this paper, we present results from a study on the interplay between head movements and phrasing and find tendencies for upward movements occuring before and downward movements occuring after prosodic boundaries.

Keywords: multimodal prosody, EMA, head movements

1 Introduction

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). Facial beat gestures align with pitch accents in speech, functioning as visual prominence markers. However, it is not yet well understood whether and how gestures and pitch accents might be combined to create different types of multimodal prominence, and how specifically visual prominence cues are used in spoken communication.

In earlier work, Ambrazaitis & House (2017) explored the patterning and usage of focal pitch accents, head beats and eyebrow beats. The material consisted of Swedish television news broadcasts and comprised audiovisual recordings of five news readers (two female, three male). They found that head beats occur more frequently in the second than in the first part of a news reading, and also that the distribution of head beats might to some degree be governed by information structure, as the text-initial clause often defines a common ground or presents the theme of the news story. The choice between focal accent, head beat and a combination of them is subject to variation which might represent a degree of freedom for the speaker to use the markers expressively.

Based on the same, but extended data, Frid et al. (2017) developed a system for detection of speech-related head movements. 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. They then used a video-based face detection procedure to extract the head positions and movements over time, and based on this they calculated velocity and acceleration features. Then a machine learning system was trained to predict absence or presence of head movement. The system achieved an F1 score of 0.69 (precision = 0.72, recall = 0.66) in 10-fold cross validation. Furthermore, the area under the ROC curve was 0.77, indicating that the system may be helpful for head movement labelling.

2 Kinematic vs audiovisual data

One difficulty in tackling the relationship between speech and the body gestures is that it requires simultaneously recorded kinematic and acoustic measurements. Previous studies have used audiovisual data to study this link, but with such data, it is not possible to compare synchronization of gestures directly. Kinematic data allow for more comparable and accurate measurements. Therefore, our current approach involves examining speech and body movements concurrently, using electromagnetic articulography (EMA). This method allows for simultaneous recording of audio + 3D movements of the articulators: tongue, lips, and jaw, but markers can also be placed on the head. Head movements are typically used to normalise, but they can also be used as raw data and thereby give us the co-occurrent position of the head. Compared to video this gives us 3D coordinates instead of 2D, and has better temporal resolution (video normally has a much lower frame rate) and better audio-video sync.

In this study we also employ it as an example of data reuse (Pasquetto et al. 2017): our material was recorded in other projects for other purposes, but we are able to use it here to study co-occurrent properties of speech and head movements. In this study we use data from one of the projects (see below).

3 Data

The data was recorded as part of the VOKART project (Schötz et al. 2013). 29 native speakers (age: 20-63) of the Stockholm (9), Gothenburg (10), and Malmö (10) variants of Swedish were recorded by means of EMA using an AG500 (Carstens Medizinelektronik) with a sampling frequency of 200 Hz. Ten sensors were attached to the lips, jaw and tongue, along with two reference sensors on the nose ridge and behind the ear to correct for head movements, using Cyano Veneer Fast dental glue. Audio was recorded using a Sony ECM-T6 electret condenser microphone.

The speech material consisted of 15–20 repetitions by each speaker of target words in carrier sentences of the type “Det va inte hV1t utan hV2t ja sa” (It was not hV1t, but hV2t I said), where V1 and V2 were different vowels. The target words containing the vowels were stressed and produced with contrastive focus. The sentences were displayed in random order on a computer screen, and the speakers were instructed to read each sentence in their own dialect at a comfortable speech rate. In order to familiarise the speakers with the sensors and the experimental setup the actual test sentences were preceded by two phonetically rich and challenging sentences, which the speakers were asked to repeat three times each. The two sentences were:

1) *Mobiltelefonen är nittiotalets stora fluga, både bland företagare och privatpersoner.* (The mobile phone is the big hit of the nineties, both among business people and private persons.)
2) *Flyget, tåget och bilbranschen tävlar om lönsamhet och folkets gunst.* (Airlines, train companies and the automobile industry are competing for profitability and people's appreciation.)

In addition, the speakers were also asked to describe a painting displayed on the computer screen, resulting in about half a minute of spontaneous speech, with several focused words and phrase-boundaries. A contour of the palate was obtained by the speakers moving their tongue tips several times back and forth along the midline of their palate. For this study, we focused on the phonetically rich sentences.

4 Analysis: head movements and phrase boundaries

We analyzed the sentence data by looking at sentence-level patterns of head movements and comparing them word by word. Sentence 1 above consists of two phrases, with an intonational boundary between the words *fluga* and *både*. Sentence 2 is essentially one phrase, but starts with a list that may cause boundary signalling. We examined the material by looking for possible head movement reflections of those boundaries. In order to get an annotation of the word boundaries of the sentences, we use the forced alignment method provided by the Praat program (Boersma & Weenink 2018), which speeds up the process but still requires manual post-checking. Utterances that contained misreadings and/or missing parts (because the recording stopped before the reader finished the sentence) were discarded. In total, there were 86 examples of Sentence 1 and 80 examples of Sentence 2.

First we measured the velocity of the angle in the sagittal plane between 1) an imaginary line between the two reference sensors (behind the ear and on the nose ridge) and 2) a line running along the transverse plane (parallel to the ground). This effectively measures the head's movement as it is tilted along this plane. We then calculated the average angular velocity per word for each sentence and then grouped the data by word. Figures 1 and 2 show summaries of the data in the form of boxplots. For Sentence 1 (in Figure 1), we note that the boundary-preceding word *fluga* has a positive median, whereas both the preceding word *stora* and the following word, *både*, has negative medians. A similar, but less prominent, pattern can be observed in Sentence 2 (Figure 2), where the first word *Flyget* has a positive median, whereas the following word *tåget*, has a negative median.

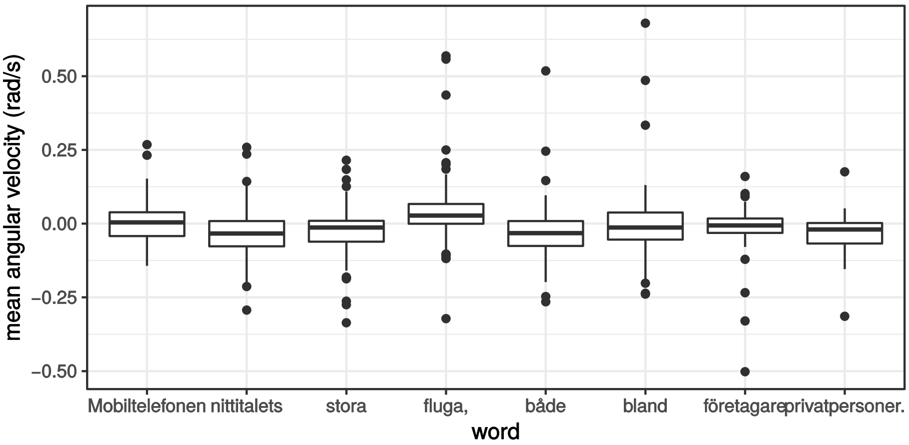

Fig. 1. Boxplots of mean angular velocity per word in sentence 1, n=86. Black horizontal lines are medians, hinges correspond to the first and third quartiles, black dots are outliers

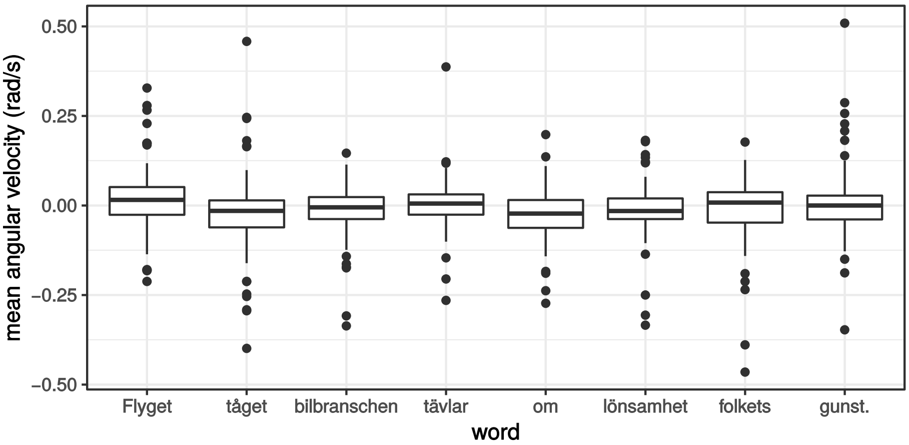

Fig. 2. Boxplots of mean angular velocity per word in sentence 2, n=80. Black horizontal lines are medians, hinges correspond to the first and third quartiles, black dots are outliers.

We used R (R Core Team, 2018) and lme4 (Bates, Maechler & Bolker, 2015) to perform a linear mixed effects analysis of the relationship between mean angular velocity and word. Linear mixed models were used to account for repeated measures. As fixed effect, we entered word into the model. As random effect, we had intercepts for subjects as well as by-subject random slopes for the effect of word. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question. P-values < 0.05 were considered significant. We tested all pairs of words w1 and w2 within each sentence, where w2 was the word following w1. Table 1 summarizes the results. The results confirm the observation that the boundary-preceding word *fluga* has a higher mean angular velocity than its neighbouring words. Furthermore, the initial word in each sentence (*Mobiltelefonen* and *Flyget*, respectively) is associated with a higher mean angular velocity, a well as the word *tävlar* compared to the word *om*.

Table 1. Results of mixed model analysis of the mean angular difference between pairs of succeeding words. Only the significantly different pairs are shown.

w1	w2	result
fluga	både	word affected mav (χ2 (1)=8.5201, p=0.003512), lowering it by about 0.077 rad/s ± 0.017 (standard errors)
stora	fluga	word affected mav (χ2 (1)=8.4946, p=0.003562), increasing it by about 0.077 rad/s ± 0.017 (standard errors)
Mobiltelefonen	nittitalet	word affected mav (χ2 (1)=5.8811, p=0.0153), lowering it by about 0.043 rad/s ± 0.012 (standard errors)
Flyget	tåget	word affected mav (χ2 (1)=3.913, p=0.04792), lowering it by about 0.043 rad/s ± 0.017 (standard errors)
tävlar	om	word affected mav (χ2 (1)=4.3803, p=0.03636), lowering it by about 0.032 rad/s ± 0.012 (standard errors)

5 Discussion/Conclusions

Using EMA recordings to analyze head movement by comparing the kinematic patterns of the sensors with the audio signal is a promising method to provide us with information on the synchronization of head movements with for example prosodic signals for prominence such as F0 excursions and syllable lengthening.

The results presented here show that there is a tendency for participants to tilt the head more upwards than downwards during the boundary-preceding words. The words which succeed the boundary, conversely show an opposite tendency indicating more downward movement. There is also a tendency for sentence-initial words to have a higher mean angular velocity than the words following them.

EMA data must be recorded on-line and obtaining it is quite laborious and less suitable for collecting large amounts. Video (AV) data is easier. However, an extension of recording EMA data is that we may augment existing AV corpora with estimated articulatory information (Ouni 2013). Since we concurrently record audio and EMA data we could build models that map acoustics to articulatory data. In this way, AV corpora may be enriched with articulatory information.

Previously motion capture data has been used to investigate temporal coordination between head movement and the audio signal (Alexanderson et al. 2013) and between head movement and EMA articulation data (Krivokapić et al. 2017; Esteve-Gilbert et al. 2018). EMA methodology has also been used to analyze head movements alone, but the current data will enable us to investigate the temporal coordination of head movements, tongue and lip movements, and the audio signal in the same system. We plan to use this methodology to investigate the role of head movement, articulation and prosody in signaling prominence in the context of the newly initiated PROGEST project and thereby contribute to the body of knowledge of multimodality in digital humanities and digital representations of speech and gestures in communication.

6 Acknowledgements

This work was supported by grants from the Swedish Research Council: Swe-Clarin (VR 2013-2003) and Progest (VR 2017-02140).

7 References

1. Alexanderson, S., House, D., & Beskow, J. (2013). Aspects of co-occurring syllables and head nods in spontane-ous dialogue. In Proc. of 12th International Conference on Auditory-Visual Speech Processing (AVSP2013). Annecy, France.
2. Ambrazaitis, G. & House, D. (2017). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings, Speech Communication, 95, pp. 100-113, https://doi.org/10.1016/j.specom.2017.08.008
3. Bates, D., Maechler M., Bolker B. & Walker S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.
4. Boersma, P. & Weenink, D. (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.39, retrieved 3 April 2018 from http://www.praat.org/
5. Esteve-Gibert, N., Loevenbruck, H., Dohen, M. & D'Imperio, M. (2018) Head movements highlight important information in speech: an EMA study with French speakers. DOI10.13140/RG.2.2.21796.78727 Conference: XIV AISV Conference - Speech in Natural Context, 25-27 January 2018, Bozen-Bolzano
6. Frid, J., Ambrazaitis, G., Svensson-Lundmark, M. & House D. (2017). Towards classification of head move-ments in audiovisual recordings of read news, Proceedings of the 4th European and 7th Nordic Symposium on Multimodal Communication (MMSYM 2016), Copenhagen, 29-30 September 2016, Volume, Issue 141, 2017-09-21, Pages 4-9, ISSN 1650-3740
7. Krivokapić, J., Tiede, M.K. & Tyrone, M. E. (2017). A Kinematic Study of Prosodic Structure in Articulatory and Manual Gestures: Results from a Novel Method of Data Collection. Lab Phonol. 2017; 8(1): 3. Published online 2017 Mar 13. doi: 10.5334/labphon.75
8. Ouni, S., Multimodal Speech: from articulatory speech to audiovisual speech. Machine Learning [cs.LG]. Université de Lorraine, 2013.
9. Pasquetto, I.V., Randles, B.M. & Borgman, C.L., (2017). On the Reuse of Scientific Data. Data Science Journal. 16, p.8. DOI: http://doi.org/10.5334/dsj-2017-008
10. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
11. Schötz, S., Frid, J., Gustafsson, L., & Löfqvist, A. (2013). Functional Data Analysis of Tongue Articulation in Palatal Vowels: Gothenburg and Malmöhus Swedish /i:, y: , ʉ:/. Proceedings of Interspeech 2013. Lyon.