EMA-based head movements, word accents, vowel length and segments: a preliminary study

Johan Frid[1], Malin Svensson Lundmark[2], Gilbert Ambrazaitis[3] and David House[4]
[1] Lund University Humanities Lab, Lund University
[2] Centre for Languages and Literature, Lund University
[3] Department of Swedish, Linnæus University
[4] Department of Speech, Music and Hearing, KTH
johan.frid@humlab.lu.se, malin.svensson_lundmark@ling.lu.se,
gilbert.ambrazaitis@lnu.se, davidh@speech.kth.se

Abstract

This study describes on-going work in the field of multimodal prosody carried out by means of simultaneous recordings of speech acoustics, articulation and head movements.

Introduction

People naturally move their heads when they speak, and head movements have been found both to correlate strongly with the pitch and amplitude of the speaker's voices and to convey linguistic information. Here, we report on a study that explores how head movement patterns vary and co-occur with lexical pitch accents (and their acoustic correlates F0 and intensity), vowel length and segmental position. The study uses data from Swedish, where there are both two lexical pitch accents and two vowel lengths that differ phonologically.

Method

We use EMA (Electromagnetic articulography), which allows for high sample rates, accurate synchronisation of kinematic and acoustic recordings, as well as three-dimensional movement data. Kinematic data is obtained by gluing small sensors on the speakers' articulators (tongue, lips, jaw). Head movement data is obtained by similar sensors on the nose ridge and behind the ears, which allows us to capture the angle of the tilt of the head. Figure 1 shows an example of nose sensor movement.

Articulatory data was collected from 18 South Swedish speakers (12 female) using a Carstens AG501. Each speaker read leading questions + sentences containing a target word from a prompter (presented eight times in random order), an arrangement employed to put a contrastive focus onto the last element in the target sentence. This left the target word in a low-prominence inducing context, hence controlling for possible effects of sentence intonation.

Material

For this study we used eight target words where pitch accent and vowel length were cross-matched so that there were two cases of each combination of word accent category and vowel length category. All words shared the similar word-initial C /m/, followed by a vowel that was either /a/ or /ɑː/. The target words were segmented and time-normalized between 0 to 1 and the head tilt angle (sagAng) was normalized for each speaker by z-transforming the angles per speaker. Spatial movements were analysed using Generalized Additive Models, which we used to test if there were effects of segmental position (C versus V in the first syllable), word accent (1 or 2) and vowel length (short or long) on sagAng. Models were fit using the maximum likelihood (ML) estimation method.

Results

Figures 2-4 show the fitted models. The Chi-Square test on the ML scores indicates that a model with the word accent distinction is significantly better than a model without it ($X^2(4.00)=632.796$, $p<2e-16$***). Similarly, a model with vowel length distinction is significantly better than a model without it ($X^2(4.00)=820.997$, $p<2e-16$***). Finally, a model with segmental position is significantly better than a model without it ($X^2(8.00)= 173.316$, $p<2e-16$***).

Discussion

The results indicate that head nod patterns that occur in synchronisation with the stressed syllable of spoken words differ with respect to word accent, vowel length and segmental position. This could possibly point to an effect of F0 and intensity on the head nod movements.

Acknowledgements

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

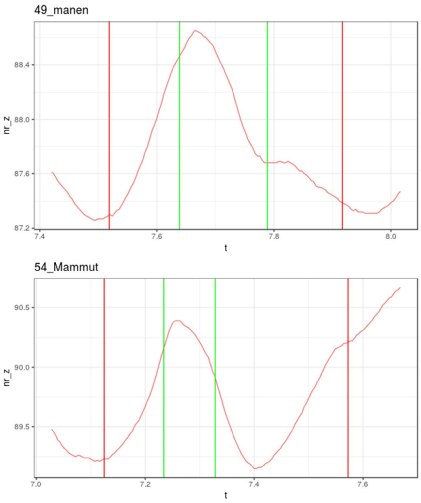

Figure 1. Two examples of nose sensor movement and alignment with vowel. CVC segment between the red lines, V between the green lines.

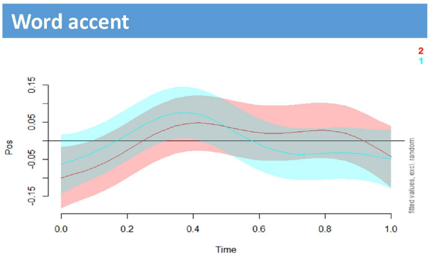

Figure 2. Non-linear smooths (fitted values) of sagAng for the Accent 1 (blue) and Accent 2 (red) words in the GAM model. Shaded bands represent the pointwise 95%-confidence interval.

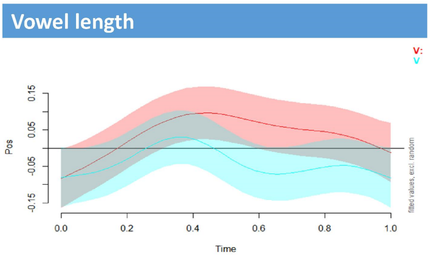

Figure 3. Non-linear smooths (fitted values) of sagAng for the V (blue) and V: (red) words in the GAM model. Shaded bands represent the pointwise 95%-confidence interval.

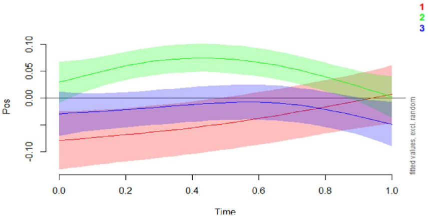

Figure 4. Non-linear smooths (fitted values) of sagAng for the pre-vocalic C (red) the V (green), and the post-vocalic C (blue) in the GAM model. Shaded bands represent the pointwise 95%-confidence interval.