CAN SEGMENTAL OR SYLLABIC DURATIONS BE PREDICTED BY THE PRESENCE OF CO-SPEECH GESTURES?

Jule Nabrotzky[1], Gilbert Ambrazaitis[1], Margaret Zellers[2], David House[3]

Linnaeus University Växjö[1], Kiel University[2], KTH Royal Institute of Technology[3]
jule.nabrotzky@lnu.se, gilbert.ambrazaitis@lnu.se, mzellers@isfas.uni-kiel.de, davidh@kth.se

ABSTRACT

Building on insights from the coordination of speech and gesture, the cumulative cue hypothesis proposes that speakers recruit cues from both modalities to signal prominence and that the relationship between the cues is additive as opposed to compensative. The study tests if durational variation is one of the cues that are recruited in this cumulative fashion in accord with gestural cues by analysing three five-minute dialogue chunks of spontaneous Swedish. The results do not indicate a direct covariation of this cue with the presence or absence of gestural cues, pointing instead to a need for further investigation of the role of duration in multimodal prominence production and perception.

Keywords: Prominence; multimodal prosody; speech-gesture integration; lexical stress; durational variation

1. INTRODUCTION

The cognitive link between speech and gestural articulation as proposed by McNeill [1] as well as Kendon [2] has been tested extensively and supported with evidence from behavioural [3] and neurological [4] domains as well as evolutionary theories [5], [6]. While the temporal coordination of speech and gesture is well-researched, the spatial coordination, that is the coordination of the magnitude and complexity of speech and gesture, is less understood. There is evidence for a spatial convergence of gesture and speech in elicited speech [7]–[9], as well as read speech in a non-experimental context [10] but there is also counterevidence [11] and the evidence for natural speech is sparse [12].

Building on the cumulative cue hypothesis [10], this paper sets out to test whether the presence of co-speech gestures or gesture clusters co-varies with auditory prominence cues. Specifically, this paper investigates the variation of segmental and syllabic durations, since this parameter has not been tested in connection with spontaneously produced gestures and in the context of spontaneous speech. To investigate this, audio-visual data taken from three dialogues of spontaneous Swedish speech are analysed regarding intonational prominence and gestures, with both manual gestures and eyebrow movements included in the analysis.

2. PREVIOUS RESEARCH

This study is built on previous research from the field of gesture studies, but also informed by knowledge about phonological features of Swedish. The literature review therefore introduces the theoretical model that generated the research question, describes relevant research on multimodal prominence, especially regarding duration and gives background on Swedish phonological characteristics that pertain to prosodic accentuation and durational variation.

2.1. Multimodal prominence

2.1.1 The cumulative cue hypothesis

Ambrazaitis and House [10] posit that speakers recruit both acoustic and visual modalities in a cumulative fashion to create stronger prominence. This means that, as in the acoustic domain, not only one method of creating prominence is used at once, but rather several acoustic and gestural characteristics increase in strength when a word receives higher prominence, which would generally be in line with Gussenhoven's [13] idea of an effort code.

2.1.2 Evidence from previous studies

Evidence for the connection of beat gestures and durations comes from a study on elicited speech in a lab where the participants were specifically asked to produce a sentence including the target words with a manual gesture, a pitch accent, both, or none [7]. The results show that longer durations are associated with both visual beats and the presence of pitch accents. In another study, the participants produced the same syllable while varying an accompanying finger tap or the emphatic stress [8]. Among other effects, the results showed lengthening in both modalities when an emphasis in one modality was present.

The cumulative cue hypothesis has also been tested with data from read speech containing spontaneously produced head and eyebrow gestures by news reporters [10]. However, while the hypothesis could be supported regarding contours of fundamental frequency, duration was not considered in the analysis. Finally, there is some evidence from a recent study on the way children highlight important information in speech [14], where the presence of head gestures was linked to longer syllable durations.

2.2. Phonological factors affecting duration

This paper investigates the potential predictive value multimodal prominence clusters (MMPs) may have on the duration of the stressed syllable. MMPs are defined by the number and type of co-speech gestures that occur together with a prosodic big accent (see 2.3). However, any variability that can be linked to MMPs will likely not be the most prominent factor in durational variation. Instead, there are several influencing factors on the segmental and word level [15]. One word level factor is the number of segments in the word and the stressed syllable. Even though stressed syllables always have the same moraic weight, the segmental complexity of the syllables can of course vary and influence the length of the individual long segments.

A factor on the segment level that is particularly important for durational variation is whether the long sound in the stressed syllable is a consonant or a vowel [16]. In Swedish, the distinction between stressed and unstressed syllables is quantity-based, with stressed syllables containing one long segment and every stressed syllable consisting of two morae. The long segment can be a vowel or a consonant, but there is an asymmetry in moraic weight between long consonants and vowels, since short vowels always have one mora, while consonants need to be lengthened to be counted as a mora [15].

2.3. Swedish word accents

Even though the f0 contour is not the object of analysis, it is important to quickly spell out the concept of Swedish word accents since they play an important role for prominence in the auditory modality.

In Swedish, each word is assigned one of two contrasting pitch accent contours. The Lund model [17] posits that both accents are two-peaked, with timing being the important distinction. The first peak in accent 1 is earlier than in accent 2 and located in the pre-stressed syllable. Additionally, the Lund model proposes a distinction between two tonal prominence levels, in which the second peaks of the accent contour is only realized in connection with sentence-level prominence [18]. The two-peaked version of each word accent is often called the "big accent", as opposed to the one-peaked "small accent" [19]. Since this paper strictly focuses of the potential accentual lengthening associated with the presence of accompanying gestures, rather than on the well-established accentual lengthening induced by the intonational prominence level (see 2.4), all words that are tested in this data are words that were realised with a big accent.

2.4 Swedish big accent lengthening

While it is well-documented that accented words are lengthened as opposed to non-accented words in stress-timed languages, the domain of that lengthening can differ between languages [20]. Additionally, Swedish accentual lengthening is better described as a big accent lengthening [16], since even words without a sentence-level prominence are assigned a word accent (as described in section 2.3).

For Swedish, experimental studies have shown that the amount of lengthening of entire words is dependent on speaker, word accent type and the position in the phrase [16], [21]. Most of the lengthening, about 75 % of the total lengthening, is confined to the stressed syllable and lengthening raises the proportion of long to short segment duration. Additionally, the consonants before and after a long vowel are lengthened, while the short vowel before a long consonant is not lengthened [16], which adds to the asymmetries between stressed consonants and vowels.

3. MATERIALS AND PROCEDURE

3.1 Materials

The materials consist of three five-minute chunks of conversation, each one between two speakers of Stockholm Swedish (in total six different speakers, three women and men each) from the Spontal corpus [22]. The recordings were filmed with two cameras and recorded with separate microphones for each speaker, making annotations sections that contain overlap between the speakers possible. The three dialogues were balanced regarding the speaker combination, with two being same-gender and one being mixed-gender.

3.2 Segmentation and annotation procedure

3.2.1 Segmental annotations

The segmental annotations of the material were done in Praat [23] and proceeded in two phases. First, all words with a big accent were annotated in accordance with [10]. In the second phase, the selected words were annotated on the syllabic and segmental level.

The syllabic annotation was conducted to find the stressed syllable and make the later annotation of onset and rhyme possible. At the segmental level, all phonemes of the word were annotated as true to the standardized pronunciation as possible, resulting in a broad annotation with the goal to be as consistent as possible while not misrepresenting the actual articulation of the speakers.

To minimize the possibility that the individual judgement of the annotators confounds the results, segmental boundaries that were unclear were set following consistent rules. Ambisyllabic consonants were always fully included in the first syllable and the boundaries between overlapping segments were annotated midway between the onset of the first and the offset of the last. If there was no consistent way the segments could be annotated (e.g., the onset of a plosive after a pause), this was marked in the data and the words were excluded in the analysis.

3.2.2 Gestural annotations

Both manual and eyebrow gestures were annotated in Elan [24] for all speakers in preparation for the analysis. Manual gestures were annotated in two phases: In the first phase, the presence and absence of manual gestures was annotated, including non-speech gestures such as scratching. This first annotation was refined through an annotation of gesture phrases in the second phase in accordance with [25]. This means that gestures were divided into preparations, strokes, retractions and holds. Beat gestures were separated into "toward" and "away" in relation to the apex and only the part of the gesture where the hand(s) moved toward the apex were included in the category of "stroke". Both hands were annotated separately but are not distinguished in the analysis, thus the presence of a gesture could mean that one hand is moving, or both.

For eyebrow movements, a fixed interval was annotated every time an eyebrow movement was visible in the video [26]. In the case of eyebrows, two researchers conducted the annotation separately and their results were compared. The annotations showed a good inter-rater reliability of κ=.698. For the eyebrow movements, only consensus annotations were included in the data.

4. RESULTS

4.1 Descriptive statistics

We identified 656 words with big accents in the material, 234 of which had to be excluded from the analysis because their segmental boundaries could not be determined in a consistent manner. 260 of the remaining 422 words were not accompanied by any gesture, while 154 were accompanied by one gesture, with manual gestures being much more common (n=123) than eyebrow gestures (n=31). Eight words in the material were accompanied by both a manual and an eyebrow gesture. The script and data files that were used in the analysis can be found in the OSF project file [27].

Before accounting for word and sentence level factors and explaining the variation in the data, a first look on the segment durations (see Fig. 1) does not reveal a clear trend. It is important to mention that both the words with eyebrow gestures and the words with manual and eyebrow gestures have a relatively low sample size.

Lexical-prosodic factors could also interact with how gestures and duration covary [10]. Fig. 2 shows the results for the duration of segments by word accent as well as multi-modal prominence cluster.

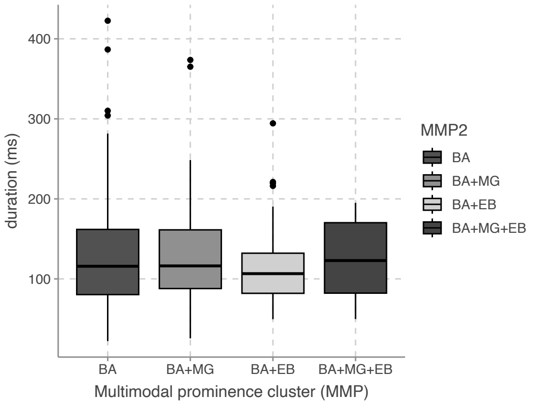

Figure 1 – The durations of weighted segments as a function of multimodal prominence clusters

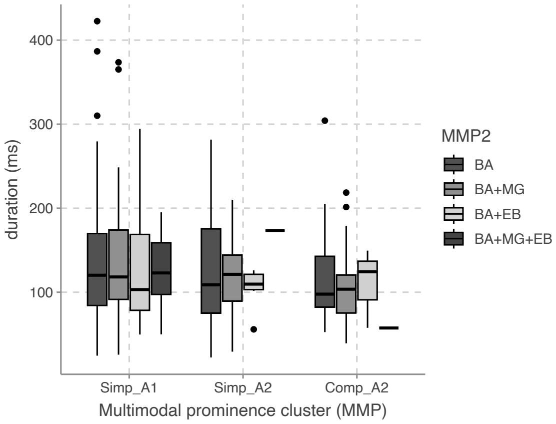

Figure 2 - Segmental durations as a function of Multimodal prominence cluster and lexical prosody

4.2 Inferential statistics

To assess the variability of duration when co-speech gestures are present, several linear mixed models were fitted to the data. Both the durations of the stressed segments, as well as the rhyme and syllable durations were tested as the dependent variables for duration. Additionally, several different combinations of predictors (multimodal prominence clusters (MMP), type of lexical prosody (LexPro) and syllable complexity (SyllCom)) for each dependent variable were tried to understand how the duration of the words in the data varied systematically. Table 1 shows the model fits with the marginal and conditional R^2-values. All models included additional fixed factors for word duration and segment type (vowel or consonant) and a random factor for speaker.

The inclusion of word duration as a fixed factor served two functions: on the one hand, it was used to account for speech rate (in combination with the random factor for speaker), and on the other hand it was also used to control for the effect that the length of the word and the number of individual segments it consists of has on the duration of each segment, which is generally believed to be inversely correlated. Segment type (consonant or vowel) was included as a fixed factor following from the moraic imbalances discussed above in sections 2.2 and 2.4.

Dep. variable	Model	R^2m	R^2c
Segment	MMP*LexPro+SyllCom	.211	.216
	MMP+LexPro+SyllCom	.208	.213
	LexPro+SyllCom	.205	.211
Rhyme	MMP*LexPro+SyllCom	.367	.367
	MMP+LexPro+SyllCom	.364	.364
	LexPro+SyllCom	.361	.361
Syllable	MMP*LexPro+SyllCom	.546	.548
	MMP+LexPro+SyllCom	.542	.544
	LexPro+SyllCom	.536	.539

Table 1 - Model fit for the segment, rhyme, and syllable level

Generally, the R^2-values indicate that the best model fits are achieved for syllable duration. However, this is to be expected since the models account for word durations. The differences between the R^2m and R^2c values indicate a low relevance of the random effect of speaker. Additionally, we conducted likelihood ratio tests to compare the full models (including the interaction term indicated by '*') with the two sets of reduced models (where either MMP is included, but no interaction is modelled, or where MMP is omitted). These tests did not reveal any significant contribution of MMP (or its interaction with LexPros).

5. DISCUSSION

The results do not support the hypothesis that prosodic prominence – here assessed in terms of accentual lengthening – and the presence of co-speech gestures covary in spontaneous Swedish speech. The likelihood ratio tests show that models that include the factor of multimodal prominence clusters do not have significantly more explanatory power. There was some variation in the results regarding which type of word accent was assigned to the word, but no significant interaction between MMP and LexPros was revealed. There are several factors that need to be addressed regarding the data.

Firstly, there is the issue of low sample sizes for subsets of data. While some subsets had a satisfactory number of tokens, the subsets of interest were rather small, owing to the relative rarity of combined gestures in the data. This might have masked effects that could have been visible with a higher number of tokens and should be considered when interpreting the results.

Another issue is that of speech rate fluctuations. In the mixed models, the word length was used to account for this instead of normalizing speech rate. However, it is not clear that this could account for all variation in speech rate and other approaches to this problem could be explored in future work.

Since these results do not correspond with previous research on this issue, it might be worthwhile to ask which of the factors that differ between the studies could have led to these results. One circumstance that also induced the problems of low sample size and speech rate fluctuations is the fact that we tested spontaneous speech as opposed to lab-elicited speech with no other confounding variables such as topic of conversation or pragmatic meaning, as well as sentence-level factors such as phrase position and rhythmic alternation that were not considered in this study. In previous lab-based investigations of gesture-speech coordination, the segmental material was held constant, simplifying the process of identifying variability in duration. Under the less controlled conditions in our experiment, building mixed models was a greater challenge, which might persist with higher sample sizes.

6. CONCLUSION

The study set out to find new evidence for the cumulative cue hypothesis that proposes the coordination of gestural and intonational prominence in a cumulative, not compensatory fashion. It focused on the aspect of durational prominence on the segmental and syllabic level. However, the data did not offer support for the cumulative cue hypothesis or any covariation of duration and gestural cues. As previously discussed, the spontaneous nature of the speech material could have been a hindrance, which points toward future research that considers a wider variety of pragmatic factors and factors on the sentence level. In addition to a better understanding durational prominence, this could lead to a more sophisticated view on the link of gesture and speech articulation on the sentence level.

7. REFERENCES

[1] D. McNeill, “So you think gestures are nonverbal?,” *Psychological Review*, vol. 92, no. 3, pp. 350–371, Jul. 1985.

[2] A. Kendon, “Semiotic diversity in utterance production and the concept of ‘language,’” *Phil. Trans. R. Soc. B*, vol. 369, p. 293, Sep. 2014.

[3] M. Graziano and M. Gullberg, “When Speech Stops, Gesture Stops: Evidence From Developmental and Crosslinguistic Comparisons,” *Front. Psychol.*, vol. 9, p. 879, Jun. 2018.

[4] R. M. Willems and P. Hagoort, “Neural evidence for the interplay between language, gesture, and action: A review,” *Brain and Language*, vol. 101, no. 3, pp. 278–289, Jun. 2007.

[5] M. C. Corballis, “From mouth to hand: Gesture, speech, and the evolution of right-handedness,” *Behav. Brain Sci.*, vol. 26, no. 02, Apr. 2003.

[6] O. Capirci and V. Volterra, “Gesture and speech: The emergence and development of a strong and changing partnership,” *GEST*, vol. 8, no. 1, pp. 22–44, May 2008.

[7] E. Krahmer and M. Swerts, “The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception,” *Journal of Memory and Language*, vol. 57, no. 3, pp. 396–414, Oct. 2007.

[8] B. Parrell, L. Goldstein, S. Lee, and D. Byrd, “Spatiotemporal coupling between speech and manual motor actions,” *Journal of Phonetics*, vol. 42, pp. 1–11, Jan. 2014.

[9] W. Pouw, L. Jonge-Hoekstra, S. J. Harrison, A. Paxton, and J. A. Dixon, “Gesture–speech physics in fluent speech and rhythmic upper limb movements,” *Ann. N.Y. Acad. Sci.*, vol. 1491, no. 1, pp. 89–105, May 2021.

[10] G. Ambrazaitis and D. House, “Probing effects of lexical prosody on speech-gesture integration in prominence production by Swedish news presenters,” *Laboratory Phonology*, vol. 24, no. 1, Aug. 2022.

[11] B. Roustan and M. Dohen, “Co-production of contrastive prosodic focus and manual gestures : temporal coordination and effects on the acoustic and articulatory correlates of focus,” in *Speech Prosody 2010,* Chicago, United States, May 2010, p. 100110:1–4.

[12] S. Berger and M. Zellers, “Multimodal Prominence Marking in Semi-Spontaneous YouTube Monologs: The Interaction of Intonation and Eyebrow Movements,” *Front. Commun.*, vol. 7, p. 903015, Jun. 2022.

[13] C. Gussenhoven, *The Phonology of Tone and Intonation*, 1st ed. Cambridge University Press, 2004.

[14] N. Esteve-Gibert, H. Lœvenbruck, M. Dohen, and M. D’Imperio, “Pre-schoolers use head gestures rather than prosodic cues to highlight important information in speech,” *Developmental Science*, vol. 25, no. 1, Jan. 2022.

[15] T. Riad, “Tonal word accents,” in *The Phonology of Swedish*, T. Riad, Ed., Oxford University Press, 2013, pp. 181–192.

[16] M. Heldner and E. Strangert, “Temporal effects of focus in Swedish,” *Journal of Phonetics*, vol. 29, no. 3, pp. 329–361, Jul. 2001.

[17] G. Bruce and E. Gårding, “A prosodic typology for Swedish dialects,” *Travaux de l’Institut de Linguistique de Lund*, vol. 13, pp. 219–228, 1978.

[18] G. Bruce, “Swedish accents in sentence perspective,” *Working papers/Lund University, Department of Linguistics and Phonetics*, vol. 12, 1977.

[19] S. Myrberg and T. Riad, “The prosodic hierarchy of Swedish,” *Nord J Linguist*, vol. 38, no. 2, pp. 115–147, Oct. 2015.

[20] T. Cambier-Langeveld and A. E. Turk, “A cross-linguistic study of accentual lengthening: Dutch vs. English,” *Journal of Phonetics*, vol. 27, no. 3, pp. 255–280, Jul. 1999.

[21] M. Heldner, “On the non-linear lengthening of focally accented Swedish words,” in *Nordic Prosody : proceedings of the VIIIth Conference, Trondheim 2000*, Frankfurt am Main: Peter Lang, 2001, pp. 103–112.

[22] J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson, and D. House, “Spontal : a Swedish spontaneous dialogue corpus of audio, video and motion capture,” in *Proc. of the Seventh LREC*, KTH, Speech Communication and Technology, 2010, pp. 2992–2995.

[23] P. Boersma and D. Weenink, “Praat : doing phonetics by computer.” Jun. 2022. Accessed: Oct. 30, 2022. [Online]. Available: praat.org.

[24] H. Sloetjes and P. Wittenburg, “Annotation by Category: ELAN and ISO DCR,” in *Proceedings of the Sixth LREC*, Marrakech, Morocco: ELRA, May 2008.

[25] A. Kendon, *Gesture Visible Action as Utterance*, 1st ed. Cambridge University Press, 2004.

[26] D. House, G. Ambrazaitis, S. Alexanderson, O. Ewald, and A. Kelterer, “Temporal organization of eyebrow beats, head beats and syllables in multimodal signaling of prominence,” in *International Conference on Multimodal Communication*, Osnabrück, 2017.

[27] J. Nabrotzky, “Can Segmental or Syllabic Durations Be Predicted by the Presence of Co-Speech Gestures?,” *OSF*, Apr. 18, 2023. osf.io/b35ke