A KINEMATIC ANALYSIS OF VISUAL PROSODY: HEAD MOVEMENTS IN HABITUAL AND LOUD SPEECH

Lena Pagel[1], Márton Sóskuthy[2], Simon Roessig[3], Doris Mücke[1]

[1] University of Cologne, [2] University of British Columbia, [3] Cornell University
lena.pagel@uni-koeln.de

ABSTRACT

Prosodic prominence manifests itself in intonation, timing and magnitude of supra-laryngeal articulation as well as speech-accompanying gestures. The interplay of prosody and gesture has been described as 'visual prosody' and is known to play an important role in communication. However, few studies have investigated visual prosody across different speaking styles. In this study, we examine co-speech head motion related to prosodic prominence in habitual and loud speech. The results show overall differences between speaking styles as well as some signatures of prosodic prominence, which are stronger in loud than in habitual speech. The paper underlines the potential of a fine-grained kinematic approach to explore continuous speech-accompanying movements.

Keywords: prosodic prominence, multimodality, co-speech gestures, loud speech, focus structure.

1. INTRODUCTION

Prosodic prominence is associated with the packaging of information in discourse, and speakers highlight the information that they deem to be most important for the listener [1], [2]. It is multi-faceted, as its correlates include a wide range of phonetic cues as well as co-speech body movements. It has been shown that prosodically prominent entities are often accompanied by certain events in the stream of co-speech gestures, such as head nods, eyebrow raises or manual gestures [3]–[7]. Since co-speech movements are tightly coupled with speech timing and serve important communicative functions, they have been termed 'visual prosody' [8].

When interacting, speakers adapt their overall speaking style to communicative demands, e.g. they may produce loud speech to be more intelligible in adverse listening conditions. Besides a number of acoustic-articulatory correlates of loud speech, it may also have an effect on speech-accompanying body movements. This is supported by evidence that co-speech head movements are interrelated with acoustic properties of speech, such as F0 and intensity [9], which are two parameters that are altered in loud speech. Additionally, studies on speech in noisy environments (Lombard speech) observed that co-speech hand, head and face gestures are enhanced in this speaking style [10]–[12]. This can have a positive effect for listeners, because co-speech gestures facilitate the perception of speech in noise [8], [13]. However, the majority of existing studies investigate Lombard speech and it is not entirely clear how these results apply to loud speech without background noise.

While loud speech can be understood as a *global increase in production effort* on the utterance level, prosodic prominence represents a *local increase* on the word or syllable level. What remains unclear to date is how these two levels interact: Do speakers use cues of visual prosody across speaking styles? The present paper explores co-speech head movements in habitual and loud speech. First, a between-style comparison assesses how head motion varies as a function of speaking style. Second, a between-focus comparison examines visual correlates of prosodic prominence in habitual and loud speech. While most studies compare prominent vs. non-prominent entities, we investigate gradient adaptations between two *degrees* of prosodic prominence (namely broad vs. corrective focus), where the words bear the nuclear pitch accent in both cases. We adopt a fine-grained kinematic approach that aims to exploit the rich information of the continuous movement signal.

2. METHODS

2.1. Recording procedure

20 native speakers of German (10 female, 10 male) between 22 and 27 years old (mean: 24.7, SD: 1.3) were recorded for this experiment. Kinematic recordings were performed using 3D Electromagnetic Articulography (EMA, AG 501) and a time-synchronised acoustic set-up. Sensors were placed behind both ears and on the bridge of the nose to calculate head movement three-dimensionally. Supra-laryngeal articulation was captured simultaneously but will be reported elsewhere.

Participants were engaged in a game-like task, in which they interacted with a virtual avatar. The game was set in a football stadium and participants were asked to answer the avatar's questions about the match, which were presented auditorily and visually. The answers could be read on the screen and were the speakers' target utterances.

2.2. Speech material

Speakers produced trisyllabic target words, which were German-sounding fictitious names with a CVCVCV structure, with the lexical stress on the penultimate syllable. They were embedded in carrier sentences and were either in *broad* or *corrective focus*. Focus structures were elicited by two kinds of questions asked by the avatar. Table 1 presents example question-answer pairs for the two focus types.

broad focus	
Q:	Was passiert gerade? 'What is going on?'
A:	[Carlotta spielt Nabima zu]$_\text{F}$. '[Carlotta passes the ball to Nabima]$_\text{F}$.'
corrective focus	
Q:	Spielt Annette Lotte zu? 'Does Annette pass the ball to Lotte?'
A:	Annette spielt [Nabima]$_\text{F}$ zu. 'Annette passes the ball to [Nabima]$_\text{F}$.'

Table 1: Example of focus elicitation. Target words are underlined, the focus domain is marked by square brackets and the subscript F.

It should be noted that the two focus types differ only in a gradient way, as target words are within the focus domain and bear the nuclear pitch accent in both conditions. However, we can observe two *degrees of prominence*, since words in corrective focus are more prominent than in broad focus [14].

All utterances were produced in *habitual* speech first, without instructions concerning the loudness level. In the second half of the experiment, speakers produced all utterances in *loud* speech. They were told that the football stadium got noisy and they had to speak up for their answers to be understood. Although no actual background noise was played, speakers increased their sound pressure level noticeably.

In total, 960 utterances were recorded (6 target words × 2 renditions × 2 focus types × 2 speaking styles × 20 speakers), out of which 954 were analysed.

2.3. Analyses

The acoustic annotation of words and segments was carried out using the Montreal Forced Aligner [15] with manual corrections. The post-processing of the kinematic data from the three head sensors was achieved using ema2wav [16]. All analyses were based on a time window that includes the target word and an additional 100 ms before and after. Since the study adopts an approach based on the continuous head motion signal that calculates kinematic and velocity profiles, no manual annotation of gestures or gesture types was carried out.

We present three separate analyses below. The first of these is based on a generalised additive mixed model (GAMM) fitted to *3D movement* trajectories for the three sensors (N = nose, R = right ear, L = left ear) across all conditions. The model includes both fixed effects for the identification of general patterns across participants as well as random smooths to capture across-participant variation. This model is used to create animations and static images that provide a schematic summary of head movements, using a triangle that represents the three sensors in 3D (cf. Figure 1). It serves the purpose of visualising the data.

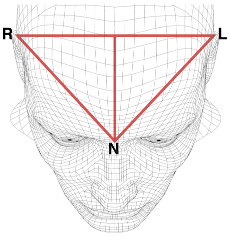

Figure 1: Visualisation of 3D movement calculation.

The remaining analyses are based on the nose sensor only, which showed the greatest degree of movement in our first measure. The second analysis captures the 3D distance between the two furthest points of each trajectory during the time window, giving an estimate of overall movement *displacement*. This measure is analysed using linear mixed effects models. The third analysis is based on an estimate of 3D *velocity* at each time point in each trajectory. This estimate is obtained by calculating the 3D distance between each adjacent time point within each trajectory in the time window, yielding velocity trajectories over time. These trajectories are analysed using GAMMs with fixed and random effects that are set up similarly to the first analysis.

The code for all analyses as well as the full data table can be found in a publicly available repository (https://osf.io/69j2s/?view_only=4ffed19ef11b4bb3b50690beab04cf73).

3. RESULTS

3.1. 3D movement visualisation

With our first measure, we use animations and static images to provide qualitative visualisations of the data. In this paper, we only include a static image taken at the temporal midpoint, but we strongly encourage readers to consult the animated versions of these visualisations in the online repository. The descriptions in the text below are based on the fuller dynamic information present in the videos.

Figure 2 shows static midpoint estimates for six representative speakers, based on the random effects of the GAMM. The majority of speakers show between-style differences, i.e. a different head position in loud speech, with the head tilted upwards (e.g., S1, S2, S3, S4). However, only a few speakers clearly differentiate focus types (e.g. S1, S3). When focus differentiation does occur (which is more often the case in loud than in habitual speech), it tends to vary in its realisation between individuals. The net effect of this variation is that, when averaging across individuals, most systematic differences are cancelled out.

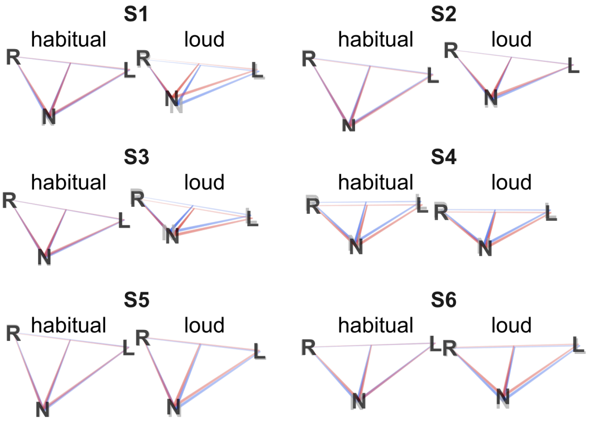

Figure 2: 3D movement results for six example speakers. Blue lines display broad, red lines corrective focus.

3.2. Displacement and velocity

The analysis of 3D displacements can provide more detailed insights (cf. Figure 3). First, a between-style analysis reveals that movements have a greater magnitude in loud than in habitual speech ($\chi^2(2) = 22.8$, $p < 0.001$, based on a model comparison). Second, the effect of focus is assessed through post-hoc comparisons with Tukey's adjustment for multiple comparisons. The results show a significant difference between focus conditions only in loud speech, with head movements being larger in corrective than in broad focus, i.e. with a higher degree of prominence ($\beta = 1.1$, $t(20) = 2.51$, $p = 0.021$).

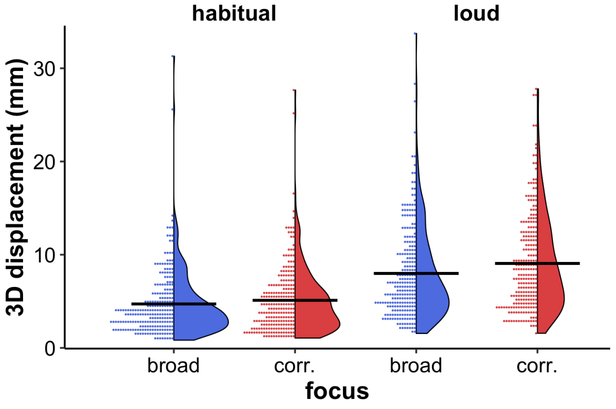

Figure 3: 3D displacement results.

Figure 4 shows results of the 3D velocity over time, based on the fixed effects from our third analysis. A between-style comparison reveals a robust difference between speaking styles: Head movements are faster in loud speech (reflected by a significant parametric difference term: $\beta = 5.32$, $t(1) = 6.97$, $p < 0.001$). A between-focus comparison shows a slight differentiation in both speaking styles: In habitual speech, focus is differentiated by the shape of the velocity profile (smooth difference term: EDF = 1, F = 6.31, $p = 0.012$) and in loud speech by the overall height of the velocity curve, i.e. faster movements are associated with a higher degree of prominence (parametric difference term: $\beta = 2.07$, $t(1) = 3.07$, $p = 0.002$).

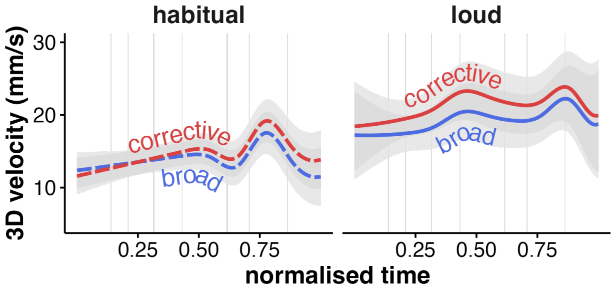

Figure 4: 3D velocity results. Grey vertical lines indicate segment boundaries of the target word.

The robust between-style differences observed at the group-level are also evident across individuals (cf. Figure 5 for six example speakers): In most speakers, there is at least a trend for head movements to be faster in loud than in habitual speech (e.g. S1, S3, S4, S6). A between-focus comparison shows that the modulations observed at the group-level are driven by few speakers who clearly differentiate focus types in at least one speaking style (e.g. S1, S4), while most speakers do not exhibit systematic differences.

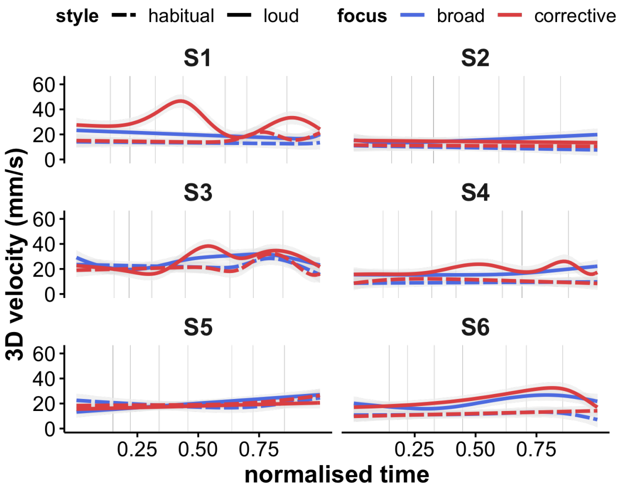

Figure 5: 3D velocity results for six example speakers. Grey vertical lines indicate segment boundaries.

4. DISCUSSION

4.1. Between-style differences

The results show reliable differences between habitual and loud speech in terms of head movements, which can be understood as correlates of global increases in production effort. In loud speech, the head is in a different position (i.e. tilted upwards) and movements are larger and faster. The majority of individual speakers exhibit these patterns of between-style differences and there is relatively little variation.

This is in line with literature showing a strong interrelation between co-speech movements and speech acoustics, namely F0 and intensity [9], which are increased in loud speech. What is more, our results are comparable with those indicating enhanced movements in Lombard speech [11], [12], [17], which underlines the similarity between the two speaking styles. The results thus show that loud speech is a multimodal phenomenon, as already described similarly for Lombard speech [17].

The head modulations in loud speech may at least partly have physiological reasons, since an upward tilt of the head (as well as changes in postural sway) can facilitate effortful speech through changes in larynx position and resonance cavities [18], [19].

4.2. Between-focus differences in both speaking styles

The results show some differentiations between focus types, or correlates of local increases in production effort. They can be interpreted as evidence for visual prosody, which is in line with existing studies [3]–[6]. Namely, co-speech head movements are larger in words associated with a higher degree of prominence, though only in loud speech. Additionally, movements exhibit different velocity profiles between focus types in both speaking styles. Interestingly, there is great inter-individual variation and the group-level results are predominantly driven by few individual speakers who strongly differentiate focus types, whereas the majority do not exhibit systematic modulations.

It can be assumed that the between-focus differences are mainly functionally and only partly physiologically motivated. This is supported by evidence that speakers are aware of communicative demands and make use of visual cues in a functional and listener-oriented way [20]. Perception experiments confirm that co-speech gestures in fact facilitate speech understanding in adverse listening conditions, showing that listeners can successfully interpret the provided multimodal cues [8].

The data show that between-focus differences are stronger in loud than in habitual speech. This suggests that speakers exploit those parameters that contribute best to the transmission of their message [21]. When the auditory channel is disturbed by (real or imaginary) background noise, the visual modality is modulated to a greater extent.

It should be noted that the two focus conditions investigated here are associated with two *degrees* of prominence that differ only in a gradient way on the level of speech production, since the target words bear the nuclear pitch accent in both focus conditions. The results of the study show that even these small differences can manifest themselves in co-speech movements, albeit only subtly, in some speakers and not always in both speaking styles. It should be noted, additionally, that we recorded co-speech movements in a highly controlled lab speech task. We hypothesise that manifestations of visual prosody may be even stronger under more natural circumstances.

4.3. Methodological proof-of-concept

The present study adopts a methodological approach analysing visual prosody in a gradient way. The kinematic analyses take all head movement during a specified time window into account and explore the rich continuous signal. The study underlines the potential of using EMA in multimodal analyses, as previously shown by, e.g., [22]. The 3D recordings can capture the multidimensional head movements occurring during speech, which may be difficult to differentiate in annotations [23].

Since speakers constantly move their heads during speech, clear onsets or targets of gestures may be challenging to detect in a traditional way, especially in the present study, where head movements are comparably small. Instead of scrutinising classifications or certain points of the gesture (e.g. the target), we consider the *movement itself* to be informative, including changes in velocity as an expression of visual biomechanical effort [24]. The methodology also enables a fine-grained exploration of individual differences that might not surface in overall results.

5. CONCLUSION

The present study investigates co-speech head movements in habitual and loud speech. The results show robust differences between speaking styles, namely enhanced movements in loud speech. They also reveal some modulations associated with prosodic prominence in both speaking styles, which are stronger in loud speech. Both findings suggest that the auditory and the visual signal can be modulated in order to meet communicative demands and convey a message. The study further underlines the potential of a multimodal analysis with 3D EMA that focusses on continuous kinematics. We conclude that speakers increase biomechanical power as a concomitant of global and local increases in production effort.

ACKNOWLEDGEMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) in the collaborative research center SFB1252 Prominence in Language (Project-ID 281511265) and project RO 6767/1-1.

REFERENCES

[1] M. A. K. Halliday, *Intonation and grammar in British English*. De Gruyter, 1967. doi: 10.1515/9783111357447.

[2] E. Vallduví and E. Engdahl, 'The linguistic realization of information packaging', *Linguistics*, vol. 34, pp. 459–519, 1996.

[3] G. Ambrazaitis and D. House, 'Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings', *Speech Commun.*, vol. 95, pp. 100–113, 2017, doi: 10.1016/j.specom.2017.08.008.

[4] J. Krivokapić, M. K. Tiede, and M. E. Tyrone, 'A Kinematic Study of Prosodic Structure in Articulatory and Manual Gestures: Results from a Novel Method of Data Collection', *Lab. Phonol.*, vol. 8, no. 1, pp. 1–26, 2017, doi: 10.5334/labphon.75.

[5] M. Swerts and E. Krahmer, 'Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions', *J. Phon.*, vol. 38, pp. 197–206, 2010, doi: 10.1016/j.wocn.2009.10.002.

[6] Y. Yasinnik, M. Renwick, and S. Shattuck-Hufnagel, 'The timing of speech-accompanying gestures with respect to prosody', *Proc. Sound Sense 11-13 June Camb. USA*, vol. 115, no. 5, pp. 2397–2397, 2004.

[7] N. Esteve-Gibert, J. Borràs-Comes, E. Asor, M. Swerts, and P. Prieto, 'The timing of head movements: The role of prosodic heads and edges', *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4727–4739, 2017, doi: 10.1121/1.4986649.

[8] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, 'Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception', *Psychol. Sci.*, vol. 15, no. 2, pp. 133–137, 2004.

[9] H. C. Yehia, 'Linking facial animation, head motion and speech acoustics', *J. Phon.*, vol. 30, pp. 555–568, 2002, doi: 10.1006/jpho.2002.0165.

[10] J. Trujillo, A. Özyürek, J. Holler, and L. Drijvers, 'Speakers exhibit a multimodal Lombard effect in noise', *Sci. Rep.*, vol. 11, p. 167211, 2021, doi: 10.1038/s41598-021-95791-0.

[11] E. Vatikiotis-Bateson, A. V. Barbosa, C. Y. Chow, M. Oberg, J. Tan, and H. C. Yehia, 'Audiovisual Lombard speech: Reconciling production and perception', in *Proceedings of the International Conference on Auditory-Visual Speech processing, 31 August - 3 September, Hilvarenbeek, The Netherlands*, 2007, pp. 45–50.

[12] M. Dohen and B. Roustan, 'Co-production of speech and pointing gestures in clear and perturbed interactive tasks: multimodal designation strategies', in *Proceedings of INTERSPEECH, 20-24 August, Stockholm, Sweden*, 2017, pp. 1–5.

[13] L. Drijvers and A. Özyürek, 'Visual Context Enhanced: The Joint Contribution of Iconic Gestures and Visible Speech to Degraded Speech Comprehension', *J. Speech Lang. Hear. Res.*, vol. 60, pp. 212–222, 2017, doi: 10.1044/2016.

[14] M. Breen, E. Fedorenko, M. Wagner, and E. Gibson, 'Acoustic correlates of information structure', *Lang. Cogn. Process.*, vol. 25, no. 7–9, pp. 1044–1098, 2010, doi: 10.1080/01690965.2010.504378.

[15] M. McAuliffe, M. Socolof, S. Mihuc, and M. Wagner, 'Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi', in *Proceedings of INTERSPEECH, 20-24 August, Stockholm, Sweden*, 2017, pp. 498–502. doi: 10.21437/INTERSPEECH.2017-1386.

[16] P. Buech, S. Roessig, L. Pagel, D. Mücke, and A. Hermes, 'ema2wav : doing articulation by Praat', in *Proceedings of INTERSPEECH, 18-22 September, Incheon, Korea*, 2022.

[18] K. Honda, 'Interactions between vowel articulation and F0 cotrol', in *Proceedings of Linguistics and Phonetics: Item Order in Language and Speech, 15-20 September 1998, Columbus, Ohio, USA*, O. Fujimura, B. D. Joseph, and B. Palek, Eds., Prague: Karolinum Press, 2000, pp. 517–527.

[19] A. Lagier, M. Vaugoyeau, A. Ghio, T. Legou, A. Giovanni, and C. Assaiante, 'Coordination between Posture and Phonation in Vocal Effort Behavior', *Folia Phoniatr. Logop.*, vol. 62, pp. 195–202, 2010, doi: 10.1159/000314264.

[20] M. Garnier, L. Ménard, and B. Alexandre, 'Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues?', *J. Acoust. Soc. Am.*, vol. 144, pp. 1059–1074, 2018, doi: 10.1121/1.5051321.

[21] M. Fitzpatrick, J. Kim, and C. Davis, 'The effect of seeing the interlocutor on auditory and visual speech production in noise', *Speech Commun.*, vol. 74, no. August, pp. 37–51, 2015, doi: 10.1016/j.specom.2015.08.001.

[22] S. Gordon, A. Vilela Barbosa, and L. Goldstein, 'Quantitative analysis of multimodal speech data', *J. Phon.*, vol. 71, pp. 268–283, 2018, doi: 10.1016/j.wocn.2018.09.007.

[23] S. Kousidis, Z. Malisz, P. Wagner, and D. Schlangen, 'Exploring Annotation of Head Gesture Forms in Spontaneous Human Interaction', *Proceedings of the Tilburg Gesture Meeting (TiGeR), 19-21 June, Tilburg, The Netherlands*. pp. 1–4, 2013.

[24] B. Lindblom, 'Explaining Phonetic Variation: A Sketch of the H&H Theory', in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds., Dordrecht: Kluwer Academic Publishers, 1990, pp. 403–439. doi: 10.1007/978-94-009-2037-8_16.