Strategies of head nod alignment with pitch prominence in French focus

Christopher Carignan[1], Núria Esteve-Gibert[2], Hélène Lœvenbruck[3],
Marion Dohen[4], Mariapaola D'Imperio[5]

[1]*Department of Speech, Hearing and Phonetic Sciences, University College London, UK*
[2]*Department of Psychology and Education Sciences, Universitat Oberta de Catalunya, Spain*
[3]*Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, LPNC, France*
[4]*Université Grenoble Alpes, CNRS, Institute of Engineering Université Grenoble Alpes (Grenoble INP), GIPSA-lab, France*
[5]*Department of Linguistics & Center for Cognitive Science, Rutgers University, USA*
c.carignan@ucl.ac.uk, nesteveg@uoc.edu, helene.loevenbruck@univ-grenoble-alpes.fr,
marion.dohen@grenoble-inp.fr, mariapaola.dimperio@rutgers.edu

Abstract

Previous research has shown that rhythmic head movement accompanies F0 modulations in speech (Munhall et al., 1994) and that this co-verbal head movement may be linked to prosodic features such as pitch accents and prosodic boundaries (Esteve-Gibert et al., 2017; Hadar et al., 1984; House et al., 2001). In this study, we examined how the production of vertical head nods may be temporally related to pitch prominence connected to different focus conditions in French. Electromagnetic articulometry data from three stable locations on the head were used to generate a time-varying signal of the changing elevation of the head. Using this signal, we examined the temporal relationship between head nod gesture strokes and F0 peaks. The results suggest that speakers use two different strategies in aligning head nods with pitch prominence, one of which mirrors alignment previously observed for oral articulatory gestures (D'Imperio et al., 2007). We also observe evidence that some speakers show a preference for one strategy over another.

Keywords: co-speech gestures, head nods, pitch prominence, electromagnetic articulometry, task dynamics

1. Introduction

Rhythmic, co-verbal movement of the head usually accompanies speech (Munhall et al., 1994). Previous work on rhythmic head gestures (head-nods and eyebrow movements) during speech specifically focused on timing and motor organization. These studies suggested that co-verbal head movements are linked to the production of prosodic features such as pitch accents and prosodic boundaries (Esteve-Gibert et al., 2017; Hadar et al., 1984; House et al., 2001), in a similar manner as co-verbal rhythmic gestures that occur with other articulators like the hands (traditionally called "beat" gestures; McNeill, 1992). Recent cross-linguistic work in Japanese and English has for instance shown that eyebrow movement tend to occur in correspondence with phrase boundaries and not prominent syllables (de la Cruz-Pavía et al., 2019; Guaïtella et al., 2009). The current study examines head movement correlates of contrastive and corrective focus in French interactive speech (e.g., 'Take the ORANGE dress [not the blue dress]'). Previous work showed that, in a similar task, French preschoolers mark focus only through head movement (but not through prosodic strategies), by accompanying contrastive and corrective focus words with more frequent head gestures than broad focus productions (Esteve-Gibert et al., n.d.). In this study we investigate whether adult speakers (who *do* use prosodic strategies) align head nods with fundamental frequency (F0) peaks, and whether the alignment is dependent on focus type (contrastive vs. corrective) or word position within the Accentual Phrase (adjective vs. noun).

2. Methodology

2.1. Task and speakers

Data presented here were collected from 12 native Southern French speakers, who participated in a game that elicited spontaneous production of sentences in three conditions (no-focus; contrastive focus; corrective focus). The spontaneous sentences were usually of the form: 'No, take the [noun] [ADJECTIVE]', and the game was designed so as to elicit two target focus positions (on the noun; on the adjective). Only data from the two focused conditions and for canonical utterances of noun + adjective sequences are included in the present study.

2.2. Identifying visually prominent head nods

Video recordings of the experimental sessions were visually inspected and annotated by the second author, using ELAN software (The Language Archive, 2015). Downward head nods that were perceived to be visually prominent were annotated, and the temporal interval of the word bearing the nod was logged for subsequent kinematic and acoustic analysis; this word will be referred to as the "target word" throughout the paper. In total, 116 visually prominent head nods were identified among the 12 speakers, ranging between 5-20 nods per speaker.

2.3. Identifying kinematically prominent head nods

During the experimental task (Section 2.1), head movement was captured using a Carstens AG500 electromagnetic articulometry (EMA) system at the Laboratoire Parole et Langage (LPL, CNRS, France). EMA data from sensors on the left and right mastoids and the nasion were captured at a sampling rate of 250 Hz. In order to estimate head orientation, a vector extending from the inter-mastoid point (i.e. the centroid between the two mastoid sensors) to the nasion was calculated, and the unit vector x-y-z components were transformed to spherical coordinates. The resultant elevation angle, ϕ, captures upward-downward angle of head movement within the spherical space defined by the inter-mastoid point as the origin and the nasion as the zenith (1, θ, ϕ). Figure 1 displays the relationship between the three head sensors and the ϕ angle. The time-varying ϕ signal was z-score normalized for each speaker and will be referred to as the "head nod signal" throughout the paper.

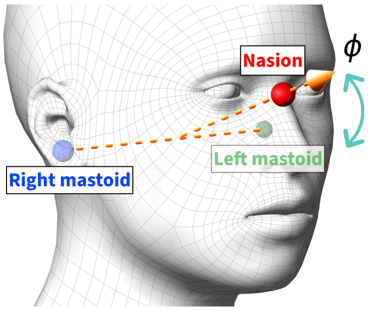

Figure 1: *Schematic representation of spherical elevation, ϕ, interpreted as vertical head angle for identifying nod gestures.*

For each utterance, the time point of maximum downward velocity of the normalized head nod signal (i.e. a downward head nod) nearest to the ELAN-annotated target word (Section 2.2) was automatically identified, and 20% velocity thresholds were used to determine the onset and the apex of the head nod (Kroos, 1996). In other words, the nod onset is defined as the point before the point of maximum velocity where velocity crosses 20% of the maximum value, and the nod apex is defined as the point after the point of maximum velocity where velocity crosses 20% of the maximum value. The interval from the onset to the apex is a phase referred to in the gesture literature as the "stroke", which is distinct from the "preparation" and "retraction" gestural phases. It is the temporal alignment of this gestural phase (the stroke) that we investigate in this study.

Figure 2 displays an example utterance with the head nod signal plotted in the solid black line and the F0 track overlaid in red circles. The acoustic interval of the noun is denoted by vertical yellow lines, the acoustic interval of the adjective is denoted by vertical blue lines, and the shared boundary between the two words is denoted by the dashed yellow-blue line. The interval of the head nod stroke is denoted by the gray rectangle. In this example, we observe that the head nod occurs primarily on the adjective *rouge* "red", but that the stroke begins before the onset of the word (i.e. within the noun *robe* "dress") and that the apex of the head nod is aligned with an F0 peak at approximately the midpoint of the adjective interval.

2.4. Identifying auditorily prominent pitch peaks

Audio recordings of the experimental sessions were inspected and annotated in Praat (Boersma & Weenink, 2020) by the third and fourth authors, who are native French speakers. The most perceptually prominent F0 peak nearest to the ELAN-annotated target word (Section 2.2) was identified in each phrase, and its time point was logged for comparison with gestural time points.

2.5. Utterance-wise time normalization

The methodological steps described in Sections 2.2-2.4 resulted in time points associated with four acoustic and articulatory events in each utterance: the F0 peak, the onset of the head nod downward movement, the point of maximum (absolute) velocity of the head nod, and the apex of the head nod. In order to compare these four time points across all 116 utterances, each time point was normalized as a percentile of the target word interval, i.e. from the start (0) to the end (100) of the target word. Thus, time points occurring prior to the word start are negative percentiles and time points occurring after the word end are percentiles greater than 100. This normalization allows for the comparison of time points relative to the target word in a way that accounts for possible differences in word duration.

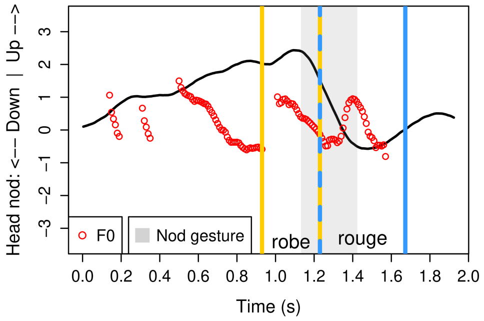

Figure 2: *Head nod signal (black, solid line) for an example utterance, with the F0 track (red circles) overlaid. The acoustic boundaries of the noun (*robe *"dress") and the adjective (*rouge *"red") are denoted by the vertical yellow/blue lines, and the interval of the head nod stroke is denoted by the gray rectangle.*

3. Results

Figure 3 displays the normalized probability distributions for the four time points: the F0 peak (red, solid line), the onset of the gesture stroke (green, dashed line), the maximum velocity of the head nod gesture (orange, dash-dotted line), and the apex of the gesture stroke (blue, dotted line). The onset and offset of the target word are denoted by the vertical black lines.

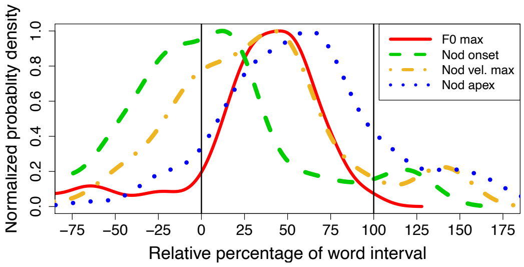

Figure 3: *Distributions of time points of F0 peaks and head nod gesture phases, relative to the target word temporal interval.*

The F0 peak in Figure 3 displays a uni-modal distribution that is aligned immediately prior to the midpoint of the target word. The head nod gesture is aligned in such a way that the point of maximum velocity is roughly aligned with the F0 peak. However, it is difficult to ascertain the precise nature of this articulatory-acoustic alignment, due to the fact that each of the head nod gesture time points displays a clear bi-modal distribution, rather than the uni-modal distribution observed for the F0 peak. Accordingly, the data were split by performing a two-group k-means clustering of the time point of maximum velocity of the head nod gesture, in order to determine the nature of the two groups underlying the global pattern observed in Figure 3. This clustering resulted in 64 items (55% of the total data) in cluster 1 and 52 items (45% of the total data) in cluster 2.

Figure 4 displays the normalized probability distributions of the data in cluster 1; we will refer to the results for this cluster as "alignment strategy 1." Here, the onset of the head nod stroke occurs before the target word, the maximum velocity of the nod is roughly aligned with the start of the target word, and the apex of the head nod stroke is aligned with the F0 peak. An example of alignment strategy 1 can seen above in Figure 2.

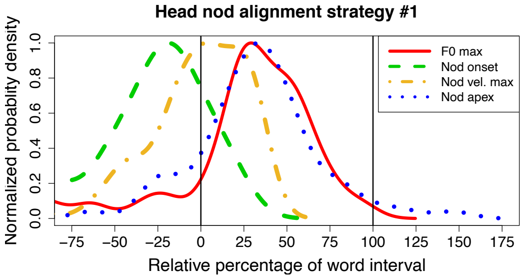

Figure 4: *Distributions of time points of F0 peaks and head nod gesture phases for alignment strategy 1 (i.e. k-means cluster 1).*

Figure 5 displays the normalized probability distributions of the data in cluster 2; we will refer to the results for this cluster as "alignment strategy 2." Here, the entire stroke of the head nod gesture is shifted forward in time compared to alignment strategy 1. In alignment strategy 2, it is the point of maximum velocity (rather than the head nod apex) that is aligned with the F0 peak. Moreover, unlike in alignment strategy 1, where the gesture stroke begins before the target word onset, the entire head nod gesture stroke occurs within the target word interval in alignment strategy 2.

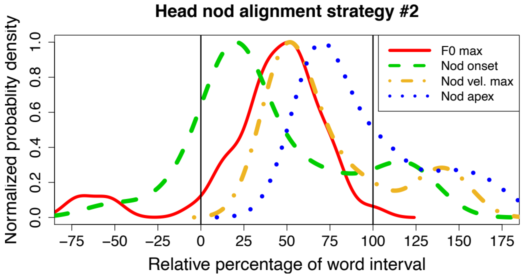

Figure 5: *Distributions of time points of F0 peaks and head nod gesture phases for alignment strategy 2 (i.e. k-means cluster 2).*

3.1. Post hoc statistical tests and exploration

In order to test for factors which may account for the differences between these two alignment strategies, a number of generalized linear mixed models (GLMMs) were constructed in R (R Core Team, 2020) using the *lme4* package (Bates et al., 2015). Each GLMM was created with logistic linking, alignment strategy as a binary dependent variable, and random intercepts and slopes by speaker. Separate GLMMs were created to test the following independent variables: focus condition (contrastive vs. corrective), target word type (noun vs. adjective), speaker-normalized F0 peak value, speaker-normalized head nod displacement (between onset and apex), speaker-normalized head nod velocity, and speaker-normalized head nod stiffness (i.e. the ratio of absolute velocity to displacement, which we use here as a proxy for gestural stiffness). No significant effects were observed except for stiffness: alignment strategy 2 was produced with greater kinematic stiffness compared to alignment strategy 1 ($p = 0.024$).

It is also useful to explore, in a qualitative manner, possible inter-speaker differences with regard to the use of these two alignment strategies. Table 1 displays the percentages of the total number of prominent head nods produced by each speaker, separated by the two alignment strategies. For ease of visualization, the magnitude of the percentage is shown both in numerical form as text and in graphical form as the strength of red color saturation (0%: white; 100%: red). The total number of prominent head nods produced by each speaker is displayed in the bottom row. Although many of the speakers produced a roughly equal number of head nods using both alignment strategies ($\leq$ 60%-40% ratio), there is some evidence to suggest that there may be speaker-specific preferences for one strategy over the other: four speakers (S01, S03, S05, S11) produced at least 70% of their head nods using alignment strategy 1, while two speakers (S02, S06) produced at least 70% of their head nods using alignment strategy 2.

4. Discussion and conclusion

We have observed in this study that downward vertical head nod gestures are temporally aligned with F0 peaks in French focus, but that there are two general strategies of alignment. In strategy 1, the onset of the gesture stroke begins before the target word interval, the point of maximum velocity is temporally aligned with the onset of the target word, and the apex of the gesture stroke is aligned with the F0 peak at roughly the temporal midpoint of the target word interval. In strategy 2, the entire head nod gesture stroke occurs within the target word interval and the point of maximum velocity of the downward movement is temporally aligned with the F0 peak at roughly the temporal midpoint of the target word. Dohen et al. (2009) also observed that inter-speaker variability in the labial correlates of focus production was associated with two main strategies: one in which the focused constituent is significantly lengthened and lip movements are hyper-articulated, and another in which the focused constituent is only slightly more articulated and the post-focus sequence is markedly hypo-articulated, thereby creating a contrast with the focused item.

Unexpectedly, the difference between the two strategies in the current study could not be accounted for either by focus type (contrastive vs. corrective) or by target word type (noun vs. adjective). The different strategies appear instead to be due to differences in kinematic control or, rather, in the specifications for the dynamical control parameters underlying the gestural movement, possibly related to hypo- vs. hyper-articulation (Dohen et al., 2009). We have observed that alignment strategy 2 was produced with significantly greater kinematic stiffness compared to alignment strategy 1. In the task-dynamics model of speech articulation, stiffer components in a second-order mechanical system move more rapidly, but also at an increased metabolic cost. Previous research has observed a decrease in oral gestural stiffness associated with unstressed or hypo-articulated speech (Perrier et al., 1996), phrase-final lengthening (Edwards et al., 1991) and, more generally, at the edges of high-level prosodic domains (Byrd et al., 2000). Thus, our findings can be interpreted in one of two ways. On the one hand, the results can be interpreted as speakers applying greater kinematic control under alignment strategy 2 in order to achieve two goals: (1) align the peak velocity of the head nod with the F0 peak, and (2) con-

Alignment Strategy	Speaker											
	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12
1	70%	29%	83%	57%	73%	29%	44%	65%	60%	50%	80%	50%
2	30%	71%	17%	43%	27%	71%	56%	35%	40%	50%	20%	50%
Total count:	10	17	6	7	11	7	16	20	5	6	5	6

Table 1: Speaker-wise percentages of head nods included in each alignment strategy group. The magnitude of the percentage is shown both in text and in (red) color saturation, and the total number of prominent head nods produced by each speaker is in the bottom row.

fine the entire head nod gesture to the temporal domain of the target word. On the other hand, the results can be interpreted as increased mechanical *compliance* under alignment strategy 1—i.e. when a gestural component of the head nod (its peak velocity) occurs at a word boundary (Byrd et al., 2000).

The fact that alignment strategy 2 emerged distinctly in these data is particularly interesting for a number of reasons. Firstly, this strategy mirrors previous findings for oral consonant articulation: accentual F0 peak targets tend to be aligned with peak velocity of the closing phase of the main oral articulatory constriction in Italian and French (D'Imperio et al., 2007). This suggests that the alignment of co-verbal head nods with pitch prominence may be part of a more general motor coordination system exploited for speech production (although, this may not be the case for upper limb movement; cf. Pouw et al., 2020). Secondly, alignment strategy 2 manifests in a speech act in which the most visually prominent event (i.e. the maximum velocity of the head nod) co-occurs temporally with the most auditorily prominent event (i.e. the F0 peak). This may suggest that speakers align different modalities in a way that maximizes the perception of prominence by the interlocutor, who then integrates the information from these different modalities in their perception (McGurk & MacDonald, 1976). In this way, head movement may be used systematically to improve auditory perception of prosodic prominence (Munhall et al., 1994).

Finally, we have observed preliminary evidence of speaker-specific strategies for one alignment strategy over the other. In some case, speakers produced both head nod alignment strategies equally. However, in other cases, some speakers produced a much larger proportion of total head nods using alignment strategy 1, while other speakers produced a larger proportion of total head nods using alignment strategy 2. Ultimately, these differences may be indicative of more general differences in the level of engagement with the experimental task (i.e. the energy level or "activation" of the speaker), as also observed in Dohen et al. (2009). Further investigation is required in order to determine the role of possible factors in conditioning these patterns of apparent speaker-specific preferences.

5. References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Soft.*, *67*(1), 1–48.

Boersma, P., & Weenink, D. (2020). Praat: Doing phonetics by computer [Software available from http://www.praat.org/].

Byrd, D., Kaun, A., Narayanan, S., & Saltzman, E. (2000). Phrasal signatures in articulation. In M. B. Broe & J. B. Pierrehumbert (Eds.), *Papers in Laboratory Phonology 5* (pp. 70–87). Cambridge Univ. Press.

de la Cruz-Pavía, I., Gervain, J., Vatikiotis-Bateson, E., & Werker, J. F. (2019). Coverbal speech gestures signal phrase boundaries: A production study of Japanese and English infant- and adult-directed speech. *Lang. Acquisition*, 1–27.

D'Imperio, M., Espesser, R., Lœvenbruck, H., Menezes, C., Nguyen, N., & Welby, P. (2007). Are tones aligned with articulatory events? Evidence from Italian and French. In J. Cole (Ed.), *Papers in Laboratory Phonology 9* (pp. 577–608). Mouton de Gruyter.

Dohen, M., Lœvenbruck, H., & Hill, H. (2009). Recognizing prosody from the lips: Is it possible to extract prosodic focus from lip features? In A. W.-C. Liew & S. Shilin (Eds.), *Visual speech recognition: Lip segmentation and mapping* (pp. 416–438). Hershey.

Edwards, J., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *J. Acoust. Soci. Am.*, *89*, 369–382.

Esteve-Gibert, N., Borràs-Comes, J., Asor, E., Swerts, M., & Prieto, P. (2017). The timing of head movements: The role of prosodic heads and edges. *J. Acoust. Soci. Am.*, *141*(6), 4727–4739.

Esteve-Gibert, N., Lœvenbruck, H., Dohen, M., & D'Imperio, M. (n.d.). Pre-schoolers use head gestures (and not prosody yet) to highlight important information in speech. *Develop. Sci.*, (under revision).

Guaïtella, I., Santi, S., Lagrue, B., & Cavé, C. (2009). Experimental investigation of the link between eyebrow movements and turn-taking. *Proc. of Gespin 2009, Poznan, Poland*, 1–6.

Hadar, U., Steiner, T. J., Grant, E. C., & Rose, F. C. (1984). The timing of shifts in head posture during conversation. *Human Mov. Sci.*, *3*, 237–245.

House, D., Beskow, J., & Granström, B. (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. *Proc. of Eurospeech 2001, Aalborg, Denmark*, 387–390.

Kroos, C. (1996). *Eingipflige und zweigipflige Vokale des Deutschen? Kinematische Analyse der Gespanntheitsopposition im Standarddeutschen* [Masters thesis, LMU Munich].

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Univ. of Chicago Press.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (1994). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychol. Sci.*, *15*(2), 133–137.

Perrier, P., Lœvenbruck, H., & Payan, Y. (1996). Control of the tongue movements in speech: The Equilibrium point Hypothesis perspective. *J. Phon.*, *24*, 53–75.

Pouw, W., Harrison, S. J., Esteve-Gibert, N., & Dixon, J. A. (2020). Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures. *J. Acoust. Soci. Am.*, *148*, 1231–1247.

R Core Team. (2020). R: A Language and Environment for Statistical Computing [Software available from http://www.R-project.org]. *R Foundation for Statistical Computing*.

The Language Archive. (2015). ELAN [Software available from https://archive.mpi.nl/tla/elan].