Head gesture timing is constrained by prosodic structure

Núria Esteve-Gibert[1], *Joan Borràs-Comes*[1], *Marc Swerts*[2], *and Pilar Prieto*[3,1]

[1]Universitat Pompeu Fabra, Spain
[2]Tilburg University, The Netherlands
[3]Institució Catalana de Recerca i Estudis Avançats

nuria.esteve@upf.edu, joan.borras@upf.edu, m.g.j.swerts@uvt.nl, pilar.prieto@upf.edu

Abstract

There is an increasing consensus to regard gesture and speech as parts of an integrated communication system, in part because of the findings related to their temporal coordination at different levels. In general, results for different types of gestures show that the most prominent part of the gesture (the apex) is typically aligned with accented syllables [6, 10-12, 14, 17]. The aim of the present study is to test for this coordination by focusing on head movements taken from a semi-spontaneous setting in order to look at the effects of upcoming phrase boundaries on their timing. Our results show that while apexes of head gestures are synchronized with accented syllables, upcoming phrase boundaries have an effect on the timing of three gestural points, namely the start, apex, and end time of head gestures. Crucially, these points are aligned differently with respect to the stressed syllable for trochees as compared with iambs/monosyllables, showing that head nods are retracted before upcoming phrase boundaries. This result corroborates previous results by Esteve-Gibert & Prieto [17] for pointing gestures in laboratory settings.

Index Terms: audiovisual speech, head gestures, prosodic structure, face-to-face communication, Catalan

1. Introduction

In face-to-face communication, meanings and intentions are conveyed by means of multimodal strategies, i.e., through both audio and visual channels. In fact, there is progressively more consensus on the idea that gesture and speech both form part of the same system of human communication [1-7]. McNeill [8] listed five main reasons to justify the tight relation between the two modalities: that they (1) co-occur in 90% of cases, (2) develop together in children, (3) are phonologically synchronous, (4) are semantically and pragmatically co-expressive, and (5) break down simultaneously in aphasia.

Several experimental studies have focused on the third of these reasons, namely that gesture and speech are synchronous from a phonological point of view. A number of these studies have found that temporal coordination can be observed between the phases of a gesture movement and related phonological events, in that the prominence in gesture and the prominence in speech co-occur in time [6, 9-14] [see 25 for an overview]. Yet, there is some debate as to what moments or events constitute the anchor points for this alignment between gesture and speech prominence. For the gesture movement, it is generally agreed that the prominence should be located in the *stroke phase*, i.e., the interval of time in which there is a peak of effort [8], or, even more precisely, at the *gesture apex*, i.e., the specific point in time (i.e., not an interval) in which the movement reaches its kinetic 'goal' [15].

There is less consensus on what constitutes the prominent part of speech that temporally coordinates with gesture. In several studies the prominent part of speech is understood as the focused word in the discourse, and they conclude that the stroke of the gesture coordinates with that word [9, 13, 16]. However, other studies take the lexically stressed syllable of (especially) that focused word as the key anchor for the gesture prominence, finding that the stroke of the gesture and the gesture apex coincide with the stressed syllable [10, 12]. And yet other studies integrate the two previous accounts and find that what aligns with the prominent part of the gesture is not simply the stressed syllable of the word in contrastive focus position, but more precisely the moment at which the pitch peak is produced within this contrastive stressed syllable [6, 11, 14, 17].

Esteve-Gibert & Prieto [17] analyzed the coordination between the gesture apex of a pointing gesture and the intonation peak in target words produced with different stress patterns (trochees, iambs, and monosyllables) by Catalan-speakers. Crucially, these words were produced in a contrastive focus condition in order to trigger different positions of the intonation peak within the stressed syllable. They found that the gesture apex and F0 peak co-occurred in time: whereas they were located at the end of the stressed syllable for trochees, they were associated with the middle of the stressed syllable for iambs and monosyllabic words. The main contribution of this article was to show that both intonation and gesture (pointing) movements were bound by prosodic phrasing, such that retracting effects occurred when there was an upcoming phrase boundary (as in monosyllables and iambs), while lagging effects occurred when there was no pre-tonic or post-tonic syllable after a preceding phrase boundary to contain part of the gesture prominence (i.e., in monosyllables).

It is important to point out that almost all the studies listed above described experimental research carried out in tightly controlled settings that hardly resemble natural interactions in face-to-face communication. Also, many of those who found that the gesture prominence coincides in time with the lexically stressed syllable analyzed deictic gestures. To our knowledge, only Loehr [18] investigated the temporal alignment between any kind of communicative hand movements (deictic gestures, iconic, and also beat gestures) and prosodic units (namely pitch accents and phrasing) in natural face-to-face interactions. Using the ToBI annotation system for American English (described in [19]), the author found that the gesture apexes coincided with pitch accents, and that the limits of the gesture phrase (defined as the combination of the gesture stroke and the preparation time needed for the gesture to reach the stroke) tended to coincide with the beginning of the intermediate phrases.

The aim of the present study is to investigate the role of two levels of prominence (accented syllables and prosodic boundaries) in the temporal coordination of head gestures and speech in semi-spontaneous face-to-face communication. Our specific purpose is to test the claim that the prosodic structure influences the timing of the gesture movement in the sense that the placement of the gesture apex within the stressed syllable depends on the metrical pattern of the target word. This has only been tested previously in laboratory settings in which participants did not have a specific communicative purpose while producing the speech and gesture signals. In the present study participants were engaged in a Guess Who game [20] and were not aware of the purpose of the study. An additional interest of the present investigation is that we focus our analysis on head gestures, which can have a potentially different behavior from other types of gestures like hand or eyebrow gestures (though some studies suggest that they show a behavior similar to that of arm and eyebrow movements; see [21] on beat gestures).

2. Methodology

2.1. Participants and procedure

Thirteen Central Catalan-speakers (1 male, 12 female), all of them undergraduates at the Universitat Pompeu Fabra in Barcelona, participated in a production task using two digital variants of the *Guess Who* board game as created by Suleman Shahid and colleagues at Tilburg University [22]. Participants played the game in pairs (i.e., with another native speaker), taking turns in adopting the different roles available. As Ahmad et al. [22] point out, the dynamic nature of games makes them a good tool for investigating human communication in different experimental setups, especially if the outcome of a game is controllable in a systematic manner.

In the *Guess Who* game, participants were presented with a board containing 24 colored drawings of human faces. These faces differed regarding various parameters, such as gender or the color of skin, hair, and eyes. Some faces were bald, some had beards or moustaches, and some were wearing hats, glasses, or earrings. As in the traditional version of *Guess Who*, the purpose of the game was to try to guess the opponent's mystery person before he or she could guess the participant's own. In this way, the game could be used to elicit in a naturalistic way target words with different metrical structures, namely trochees (e.g., *DOna* 'woman', *BARba* 'beard', *NEgre* 'black'), as well as iambs (e.g., *marRONS* 'brown', *barRET* 'hat', *verMELL* 'red') and monosyllables (e.g., *ROS* 'blond', BLAUS 'blue', NOI 'boy')[1].

During the game, participant A had to ask participant B questions to try to determine the mystery person on B's board. Players took turns asking questions about the physical features of their respective "mystery persons" in an effort to eliminate the wrong candidates. The winner was the player who guessed his/her mystery person first. In order to elicit not only questions but also statements, a variation of the game was designed. In this statement-elicitation variation of the game, participants took turns making statements about their mystery person, while the other player listened and eliminated all characters that did not exhibit a particular feature. Again, it was the player who guessed the identity of their "mystery person" first that won. Both participants within a pair took turns in the course of both variations of the game and therefore both provided examples of questions and statements.

Crucially for our goals, the types of simple questions and statements elicited with this procedure had the target words in focus position at the end of the prosodic phrase (e.g., *És una dona*? 'Is it a woman?', *És un home* 'It's a man', *Té bigoti?* 'Has he got a moustache?', *Porta un barret* 'She wears a hat', etc.).

Participants sat in the same room, facing each other across a table and in front of two laptop computers arranged so that they could not see each other's screen. Two camcorders were placed in such a way that they could record the upper part of each participant's body. Once the participants were seated, the camera was raised or lowered according to the participant's height. The experimenter then explained the game and gave instructions about the procedures to be followed for each of the two variations, which took place consecutively. Altogether each game lasted approximately twenty minutes, the time it took to play and win both variants in each set.

2.2. Coding

The relevant utterances (i.e., the questions about the mystery person and the statements used as cues) were annotated in terms of speech and gesture features. For speech, we used Praat [23] to mark the beginning of the opponent's responses and to indicate the duration of the word in focus position as well as the nuclear syllable. Then we imported the Praat label files into ELAN [24]. Figure1 shows an example of labeling with ELAN.

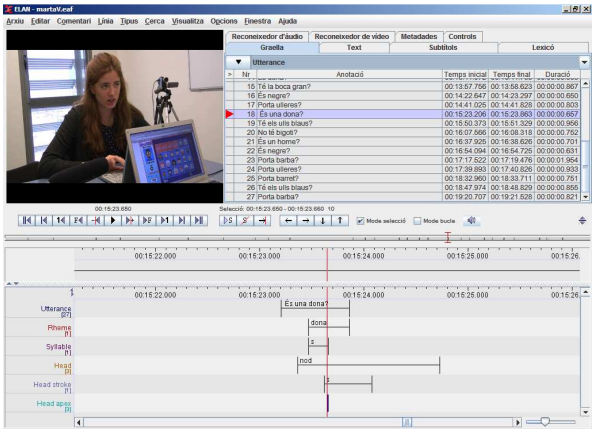

Figure 1: *Example of ELAN labeling.*

As for the gesture annotation, three tiers were created in ELAN, one to label the temporal limits of the head movement, another to locate the stroke of the head movement, and another to label the location of the gesture apex. Four possible head movement were taken into account: head nod, head upward, head tilt and other. *Head nod* referred to a downwards confirmation movement of the head; *upward* referred to a head movement directed upwards (in the opposite direction from

[1] Capital letters indicate the stressed syllable.

nodding); *head tilt* referred to a head inclination or sidewards movement; and *other* referred to any other movements timed with speech, e.g. negation gestures. Following the standard procedure, the annotation of the head movement timing consisted of locating the three gesture phases, namely the preparation phase, the gesture stroke, and the retraction phase. The head movement apex was located at the peak of effort of the head movement [4, 8].

3. Results

The total number of head gestures annotated was 114, consisting of 53 head nods, 42 head tilts, 15 head upwards, and 4 head gestures labeled 'other'. 67 of these gestures appeared in statements and 47 in questions.

3.1. Timing of the gesture apex

Figure 2 shows the temporal distance between the head gesture apex with respect to the end of the accented syllable. These results show that the gesture apex is aligned approximately with the end of the syllable for trochees (M = –67 ms), while it is aligned earlier in the case of monosyllables (M = –265 ms) and iambs (M = –404 ms). These results are consistent with the tendencies described in the literature for stress-final words.

A one-way ANOVA was run with the distance between the gesture apex and the stressed syllable end in milliseconds as the dependent variable and the stress pattern (three levels: trochees, monosyllables, and iambs) as the independent variable. Stress pattern was found to be significant ($F(2, 111) = 5.72$, $p = .004$, ${}_p\eta^2 = .09$). Bonferroni post-hoc tests revealed significant differences between trochees and monosyllables ($p < .05$), and also between trochees and iambs ($p < .05$), but not between monosyllables and iambs (p = n.s.).

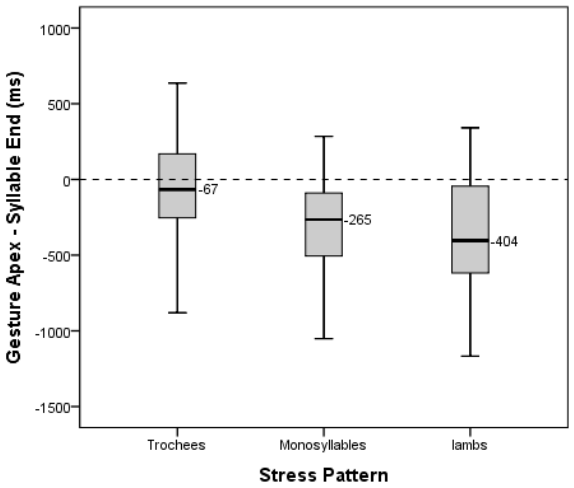

Figure 2: *Distance in time between the apex of the head gesture and the end of the accented syllable (in ms), as a function of the stress pattern of the word (trochees, monosyllables, and iambs).*

The results in this section reveal that the temporal location of the apex of head gestures is significantly affected by the distance to the upcoming phrase boundaries, that is, the apex has to be retracted when the gesture associates with word-final nuclear syllables. Interestingly, this replicates [17]'s results for the coordination between pointing gestures and speech, as they found that the apex of the pointing gesture was retracted before an adjacent phrase boundary (as in monosyllables and iambs), in comparison with gestures associated with non-adjacent phrase boundaries (as in trochees).

3.2. Timing of the gesture start

Figure 3 shows the temporal distance between the start of the head gesture with respect to the end of the accented syllable. These results show that for trochees the head gesture start is aligned closer relative to the end of the syllable (M = –389 ms), than in the case of monosyllables (M = –670 ms) and iambs (M = –734 ms), where it is aligned much earlier.

A one-way ANOVA was run with the distance between the gesture start and stressed syllable end in milliseconds as the dependent variable and stress pattern as the independent variable. Stress pattern was found to be significant ($F(2, 111) = 7.30$, $p = .001$, ${}_p\eta^2 = .12$). Bonferroni post-hoc tests revealed significant differences between trochees and monosyllables ($p < .01$), and also between trochees and iambs ($p < .05$), but not between monosyllables and iambs (p = n.s.).

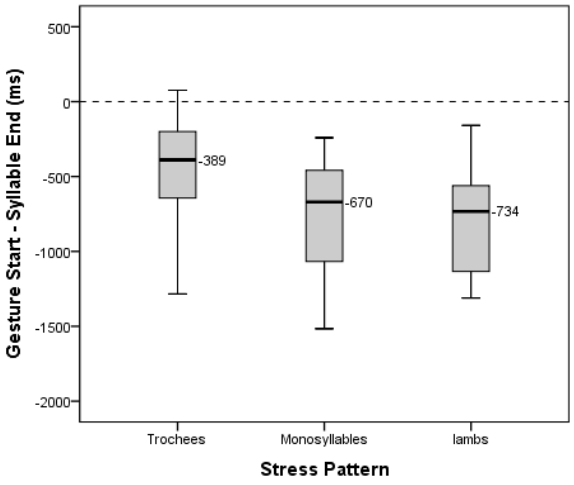

Figure 3: *Distance in time between the start of the head gesture and the end of the accented syllable (in ms), as a function of the stress pattern of the word (trochees, monosyllables, and iambs).*

The results in this section again show an asymmetry between the temporal association of head gestures with trochaic words as compared to iambic and monosyllabic words. In our interpretation, the fact that head gestures start earlier in monosyllables and iambs with respect of the end of the accented syllable indicates that the scope of the head movement is the entire focused word, not only the accented syllable, although results in 3.1 indicate that the anchoring landmark in speech for the gesture apex is the accented syllable.

3.3. Timing of the gesture end

Figure 4 shows the distance in time between the gesture end and the end of the accented syllable. These results show that the gesture end is more closely aligned with the end of the syllable in the case of monosyllables (M = 111 ms) and iambs (M = –21 ms), than in the case of trochees (M = 344 ms).

A one-way ANOVA was run with the distance in time between the gesture end and the stressed syllable end as the dependent variable and stress pattern as the independent variable. Stress pattern was found to be significant ($F(2, 103) = 5.44$, $p = .006$, ${}_p\eta^2 = .10$). Bonferroni post-hoc tests revealed significant differences between trochees and monosyllables ($p < .05$), and also between trochees and iambs ($p < .05$), but not between monosyllables and iambs (p = n.s.).

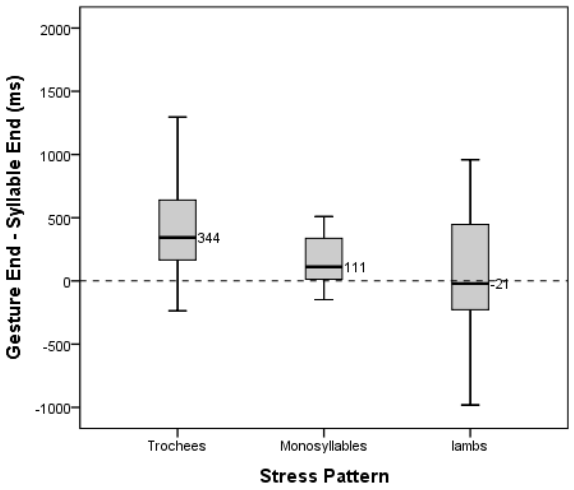

Figure 4: *Distance in time between the gesture end and the end of the accented syllable (in ms), as a function of the stress pattern of the word (trochees, monosyllables, and iambs).*

The results of this section show that the end time of head gestures is located later in trochees than in monosyllables and iambs, something that is related to the fact that this stress pattern has a final unstressed syllable available that can accommodate the retraction phase of the head gesture.

4. Discussion and conclusions

The aim of the present study was to test the claim that prosodic structure influences the timing of gesture movements in the sense that the location of the different phases of the gesture with respect to the stressed syllable depends on the metrical pattern of the target word. This has been tested with head gestures observed in a semi-spontaneous setting, while previous studies examined co-speech gestures produced in laboratory controlled settings.

Our results showed that head gestures are aligned differently with respect to the stressed syllable for trochees than they are for iambs and monosyllables. In trochees, the apex of the gesture is closely aligned with the end of the stressed syllable, the gesture start occurs together with the start of the stressed syllable, and the end of the gesture is located within the final unstressed syllable. By contrast, in iambs and monosyllables, the apex is located in the middle of the stressed syllable, the start of the head movement occurs well before the accented syllable and the ending time occurs right after the accented syllable. These results reveal that the scope of the entire head gesture movement operates on the entire focused word, since preceding and upcoming word boundaries determine the start and end time of the gesture. However, the timing of the gesture prominence (the gesture apex) is determined by the position of the stressed syllable: it occurs earlier when the stressed syllable is followed by a phrase boundary, and it occurs later with respect to the end of the stressed syllable when there is post-tonic material where to accommodate the retraction phase of the gesture. All in all our results show that the timing patterns of head gestures are constrained by prosodic structure and corroborate previous findings in the sense that the most prominent part of the head gesture (the apex) has been shown to be aligned with accented syllables [6, 10-12, 14, 17].

Further research is needed to investigate potential effects of sentence type on the temporal coordination of gesture and prosody. The prosodic structure of statement and questions is different, and this might have an impact on the temporal coordination. Our study does not have enough data to undertake these comparisons, so future studies are crucial to shed light on this issue. The study of the temporal coordination between gesture and speech (in particular, prosody) is important to understand how both modalities are entrained and if they are in fact part of the same system in communication, as proposed in the literature [1, 4, 8]. The fact that gesture and speech are produced in different physiological systems might impose biomechanical constraints in their coordination [25]. However, the evidence presented in this study and in previous work suggests that the analysis of the temporal coordination of both modalities is crucial to investigate the cognitive processes involved in speech and gesture production.

5. Acknowledgments

We would like to thank Suleman Shahid and Constantijn Kaland for their help with setting up the recording sessions, and Igor Jauk for his help with labeling the Catalan corpus. We are grateful to the participants in all the experiments for voluntarily giving us their time. This research has been funded by a research grant awarded by the Spanish Ministry of Science and Innovation (BFU2012-31995 “Gestures, prosody and linguistic structure”), by a grant awarded by the Generalitat de Catalunya (2009SGR-701) to the *Grup d’Estudis de Prosòdia*, and by a “Study abroad scholarship for research outside of Catalunya” 2010 BE1 00207, awarded by the Generalitat de Catalunya.

6. References

[1] Birdwhistell, R. L., “Introduction to kinesics: An annotated system for analysis of body motion and gesture”. Department of State, Foreign Service Institute, Washington DC, 1952.

[2] Birdwhistell, R. L., “Kinesics and context: Essays on body motion communication”, University of Pennsylvania Press, Philadelphia, 1970.

[3] Kendon, A., “Some relations between body motion and speech. An analysis of an example”, in W. Siegman & B. Pope (Eds.), *Studies in dyadic communication*, Pergamon Press, New York, NY,177-210, 1972.

[4] Kendon, A., “Gesticulation and speech: Two aspects of the process of utterance”, in M. R. Key (Ed.), *The relationship of verbal and nonverbal communication*, Mouton, The Hague, The Netherlands, 207-227, 1980.

[5] Kita, S., “How representational gestures help speaking”, in D. McNeill (Ed.), *Language and Gesture*, Cambridge University Press, Cambridge, 2000.

[6] De Ruiter, J. P., *Gesture and speech production*, unpublished doctoral dissertation). Katholieke Universiteit, Nijmegen, The Netherlands, 1998.

[7] McNeill, D., *Language and Gesture: Window into Thought and Action*, Cambridge University Press, Cambridge, 2000.

[8] McNeill, D., *Hand and mind*, University of Chicago Press, Chicago, IL, 1992.

[9] Butterworth, B. and Beattie, G., “Gesture and silence as indicators of planning in speech”, in R. Campbell & G. T. Smith (Eds.), *Recent advances in the psychology of language: Formal and experimental approaches*, Plenum Press, New York, NY, 347-360, 1978.

[10] Loehr, D. P., “Aspects of rhythm in gesture and speech”, *Gesture*, 7: 179-214, 2007.

[11] Nobe, S., *Representational gestures, cognitive rhythms, and acoustic aspects of speech: A network/ threshold model of gesture production*, unpublished doctoral dissertation, University of Chicago, 1996.

[12] Rochet-Capellan, A., Laboissière, R., Galván, A., and Schwartz, J. L., “The speech focus position effect on jaw–finger coordination in a pointing task”, *Journal of Speech, Language, and Hearing Research*, 51: 1507-1521, 2008.

[13] Roustan, B. and Dohen, M., “Co-production of contrastive prosodic focus and manual gestures: Temporal coordination and effects on the acoustic and articulatory correlates of focus”, *Proceedings of Speech Prosody 2010 Conference*, 100110, 1-4, 2010.

[14] Rusiewicz, H. L., Shaiman, S., Iverson, J. M., and Szuminsky, N., “Effects of perturbation and prosody on the coordination of speech and gesture”, *Speech Communication*, 57,283-300, 2014.

[15] Kendon, A., *Gesture: Visible action as utterance*, Cambridge University Press, Cambridge, 2004.

[16] Ferré, G., “Timing relationships between speech and co-verbal gestures in spontaneous French”, *Proceedings of Language Resources and Evaluation, Workshop on Multimodal Corpora*, 86-91, 2010.

[17] Esteve-Gibert, N. and Prieto, P., “Prosodic structure shapes the temporal realization of intonation and manual gesture movements”, *Journal of Speech, Language, and Hearing Research* 56: 850-864, 2013.

[18] Loehr, D., “Temporal, structural, and pragmatic synchrony between intonation and gesture”, *Laboratory Phonology*, 3(1): 71-89, 2012.

[19] Beckman, M. and Ayers-Elam, G., “Guidelines for ToBI labeling”, ver. 3, Ohio State University, 1997.

[20] Borràs-Comes, J., Kaland, C., Prieto, P., and Swerts, M., “Audiovisual Correlates of Interrogativity: A Comparative Analysis of Catalan and Dutch”, *Journal of Nonverbal Behavior*, in press, published online: 05 October 2013.

[21] Krahmer, E. and Swerts, M., “The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception”, Journal of Memory and Language, 57(3): 396-414, 2007.

[22] Ahmad, M. I., Tariq, H., Saeed, M., Shahid, S., and Krahmer, E., “Guess who? An interactive and entertaining game-like platform for investigating human emotions”, in Jacko, J. A. (Ed.), *Human-computer interaction. Towards mobile and intelligent interaction environments*, vol. 3, *Lecture Notes in Computer Science*, 6763, Springer, Berlin, 543-551, 2011.

[23] Boersma, P. and Weenink, D., *Praat: Doing phonetics by computer*, version 5.3.04, computer program, 2012.

[24] Lausberg, H. and Sloetjes, H., “Coding gestural behavior with the NEUROGES-ELAN system”, *Behavior Research Methods, Instruments, & Computers*, 41: 841-849, 2009.

[25] Wagner, P., Malisz, Z., and Kopp, S. “Gesture and speech in interaction: An overview”. Sp. Comm 57: 209-232, 2014.