The timing of head movements: The role of prosodic heads and edges

Núria Esteve-Gibert
Aix Marseille Université, CNRS, Laboratoire Parole et Langage, Aix-en-Provence, France

Joan Borràs-Comes[a)]
Universitat Autònoma de Barcelona, Bellaterra, Spain

Eli Asor
Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain

Marc Swerts
Department of Communication and Information Sciences, Tilburg University, Tilburg, the Netherlands

Pilar Prieto[b)]
ICREA - Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain

(Received 23 September 2016; revised 23 March 2017; accepted 4 June 2017; published online xx xx xxxx)

This study examines the influence of the position of prosodic heads (accented syllables) and prosodic edges (prosodic word and intonational phrase boundaries) on the timing of head movements. Gesture movements and prosodic events tend to be temporally aligned in the discourse, the most prominent part of gestures typically being aligned with prosodically prominent syllables in speech. However, little is known about the impact of the position of intonational phrase boundaries on gesture-speech alignment patterns. Twenty-four Catalan speakers produced spontaneous (experiment 1) and semi-spontaneous head gestures with a confirmatory function (experiment 2), along with phrase-final focused words in different prosodic conditions (stress-initial, stress-medial, and stress-final). Results showed (a) that the scope of head movements is the associated focused prosodic word, (b) that the left edge of the focused prosodic word determines where the interval of gesture prominence starts, and (c) that the speech-anchoring site for the gesture peak (or apex) depends both on the location of the accented syllable and the distance to the upcoming intonational phrase boundary. These results demonstrate that prosodic heads and edges have an impact on the timing of head movements, and therefore that prosodic structure plays a central role in the timing of co-speech gestures.
 [http://dx.doi.org/10.1121/1.4986649]

[BVT] Pages: 1–13

I. INTRODUCTION

Studies in the last few decades have shown that co-speech gestures are closely linked to speech in several ways. First, gestures and speech align in terms of semantic and pragmatic meaning (e.g., Bergmann *et al.*, 2014; Kelly *et al.*, 2010; Özyürek *et al.*, 2007). If you tell your friend that you just called your sister, it could well be that you produce a concomitant "calling" gesture in a way that the gesture represents what you also say in speech. Second, gesture and speech co-occur together, they are temporally aligned (e.g., Kendon, 1980; McNeill, 1992). When we speak, the timing of our gestures is not random but is determined by the accompanying speech. In this study, we will examine in detail the temporal alignment patterns between head gestures and speech.

Kendon (1980) and McNeill (1992) stated that the central part of a gesture movement tends to occur within the limits of the prominent prosodic elements of the speech stream. Depending on the gesture and the way it is produced, this prominent part of the gesture can be either an interval, called "gesture stroke," or a peak in the gesture movement, called "gesture apex." Many studies have further investigated the specifics of this temporal alignment, revealing that gesture strokes and gesture apexes are aligned with stressed syllables in the speech stream (see Wagner *et al.*, 2014, for a complete review). Interestingly, certain stressed syllables seem to attract more strongly the presence of co-speech gestures: gesture apexes (the peak of prominence in a gesture movement) are more frequently aligned with pitch-accented syllables and with focal pitch accents than with stressed syllables that have a lesser degree of prosodic emphasis (e.g., Alexanderson *et al.*, 2013; De Ruiter, 1998; Ferré, 2014; Yasinnik *et al.*, 2004).

Gesture-speech temporal patterns have been analysed in several contexts, from spontaneous conversations (e.g., Jannedy and Mendoza-Denton, 2005; Loehr, 2012; Yasinnik

[a)]Also at Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain.
[b)]Electronic mail: pilar.prieto@upf.edu

et al., 2004) to controlled laboratory settings (e.g., De Ruiter, 1998; Esteve-Gibert and Prieto, 2013; Leonard and Cummins, 2011; Rochet-Capellan *et al.*, 2008; Rusiewicz *et al.*, 2013). Manual gestures are by far the most-studied gestures, beat and pointing manual movements traditionally receiving most of the researchers' attention (e.g., Kendon, 1980; Leonard and Cummins, 2011; Treffner *et al.*, 2008, for beat gestures; De Ruiter, 1998; Levelt *et al.*, 1985; Rochet-Capellan *et al.*, 2008; Roustan and Dohen, 2010, for pointing gestures). Leonard and Cummins (2011) used a motion caption system to track hand gestures while participants were reading a short fable. The authors correlated five movement points (the onset of the movement, the peak velocity of the extension phase, the point of maximum extension of the hand before retraction, the peak velocity of the retraction phase, and the termination of the gesture) with three speech landmarks (the vowel onset of the stressed syllable in each word, the estimated P-centre, and the pitch peak within the stressed syllable). They found that the point of maximum arm extension (the apex) occurred while the speaker produced the stressed syllable, and that this pattern was very stable, meaning that this was the gesture landmark that showed less variability with respect to its speech anchoring.

Yet, another prosodic event might be influencing gesture timing as well, i.e., intonational phrase boundaries. There is evidence that the scope of gestural movements typically finishes at the end of intonational phrases (Loehr, 2012; Shattuck-Hufnagel *et al.*, 2010; see Krivokapić, 2014, for a review) and that listeners can automatically extract prosodic structure by using the temporal scope of manual beat gestures and thus use these gestural features disambiguating the syntactic structure (Guellaï *et al.*, 2014). Interestingly, phrase boundaries seem to impact not only the ending point of a gesture movement, but also the timing of the distinct gesture phases in relation to speech landmarks (De Ruiter, 1998; Esteve-Gibert and Prieto, 2013; Krivokapić *et al.*, 2015; Krivokapić *et al.*, 2016; Levelt *et al.*, 1985). Esteve-Gibert and Prieto (2013) observed that the movement pattern of the manual pointing gestures mimicked that of F0 movements. That is, both gesture peaks of pointing gestures and F0 peaks in rising pitch accents were retracted when the accented syllable was in phrase-final position; by contrast, they occurred at the end of the accented syllable when this syllable was non-phrase-final. Interestingly, Krivokapić *et al.* (2015) controlled the level of prosodic phrasing (no boundary, prosodic word, intermediate phrase, intonational phrase) and of prominence (de-accented, broad focus, narrow focus, contrastive focus) to see how these patterns affected the alignment of oral and manual pointing gestures with speech. The authors measured the duration of closing and opening oral movements and the duration of launching (the distance between the beginning of the pointing and its apex) and retraction (the distance between the apex and the end of the pointing) phases of the pointing gesture. The results showed that the pattern of manual gestures was very similar to that of oral gestures: oral movements were longer in trials with stronger phrase boundaries (just like the launching part of pointing gestures was), and oral movements were also longer under prominence (just like the retraction part of the pointing gestures was).

Motion caption systems have been used to explore the timing of head gestures with the aim of creating virtual agents that can engage in synthesized dialogues that are as natural as possible. These studies take the position of the accented syllables as the key prosodic landmark with which gesture movements align, but they do not take into account intonational phrase boundaries. In general, they found a similar temporal alignment pattern as had been shown for hand gestures: accented syllables are the anchoring point in speech for the most prominent part of a head movement, the gesture apex (defined as the specific point in time when the head changes its direction in the vertical or lateral movement) (Alexanderson *et al.*, 2013; Ambrazaitis *et al.*, 2015; Fernández-Baena *et al.*, 2014; Goldenberg *et al.*, 2014; Graf *et al.*, 2002; Hadar *et al.*, 1983; Ishi *et al.*, 2014; Kim *et al.*, 2014). However, these studies also reported variability in this alignment pattern. Alexanderson *et al.* (2013), for instance, analysed 54 head nods that co-occurred with target words in 20 min of spontaneous conversations, and found that the head gesture apexes occurred within the accented syllable, but that there was a great temporal variability in the precise anchoring point of the gesture apexes within that syllable. We hypothesize that this variability can be partly explained by the effects of upcoming intonational phrase boundaries.

The present study aims at investigating the role of the position of prosodic heads (accented syllables) and prosodic edges (prosodic word boundaries and intonational phrase boundaries) on the timing of head nod gestures. To our knowledge, only three studies have previously alluded at the combined effect of prosodic heads and edges but without testing it in a systematic way. Ishi *et al.* (2014) found that, in Japanese, head nods co-occur with the phrase-final syllables that are immediately followed by strong intonational phrase boundaries. Barkhuysen *et al.* (2008) observed that speakers use the visual information of head movements together with acoustic cues to mark the ends of utterances. Finally, Hadar *et al.* (1983) observed that some head gestures were associated with stress and with junctures (ends of phrases). None of these previous studies on head nod timing, however, controlled the potential effect of the position of intonational phrase boundaries on the timing of head nod movements. In our study, we want to contribute to the previous literature by adding this factor to our analysis. On the one side, we hypothesize that accented syllables (prosodic heads) attract the peak of head movements (the gesture apex). On the other side, we hypothesize that the role of prosodic edges is crucial in determining the precise location of the head apex within the accented syllable. This would imply that speakers plan the timing of their co-speech gestures by taking into account the specific characteristics of the prosodic units of speech they are associating the gesture with, and, importantly, they take into account both its prominent bits and its ending edges. If this is the case, our results would help clarifying the nature of the temporal alignment between head movements and speech events.

To investigate these hypotheses, two experiments were designed. Experiment 1 elicited spontaneous head movements that co-occurred with end-of-utterance target words displaying different stress patterns (stress-initial, stress-final, stress-medial, or monosyllables). This enabled us to test how different positions of the accented syllable and of the phrase boundary influence the timing of head movements. Experiment 2 sought to confirm the findings from experiment 1 in a more controlled way by (a) narrowing down the pragmatic function of head gestures (e.g., a confirmatory function), (b) analysing a balanced number of cases per condition, and (c) varying systematically the position of prosodic heads and edges.

II. EXPERIMENT 1

Experiment 1 examines the influence of the position of accented syllables and intonational phrases boundaries on the timing of head gestures that co-occur with spontaneous speech.

A. Method

1. Participants

Thirteen Catalan speakers (1 male and 12 females), between 19 and 24 years of age (mean age 20.9 years) participated in the experiment. All of them were undergraduates at the Universitat Pompeu Fabra in Barcelona, Spain. The participants signed a consent form and received 5 Euro as monetary compensation.

2. Materials

Two digital variants of the Guess Who board game were presented (Ahmad *et al.*, 2011), each containing 24 coloured drawings of human faces. These faces differed regarding various parameters, such as gender or the colour 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.

The game was designed to naturally elicit sentences containing target words that had different metrical patterns and different distances to upcoming intonational phrase boundaries: stress-initial words (or strong-weak words, hereafter SW) such as *dona* "woman" or *barba* "beard," stress-final words (or weak-strong words, hereafter WS) such as *marrons* "brown" or *barret* "hat," monosyllables (hereafter S) such as *ros* "blond" or *blau* "blue," and stress-medial words (or weak-strong-weak words, hereafter WSW) such as *bigoti* "mustache" or *ulleres* "glasses." These patterns displayed variability in terms of the position of the accented syllable within the prosodic word and also in terms of the distance of the accented syllable from an upcoming intonational phrase boundary. More specifically, while in the WS and S words, the accented syllables were adjacent to the right-edge intonational phrase boundary, in the SW and WSW words, there was one unaccented syllable preceding the upcoming phrase boundary.

Two variants of the game were created, a question-eliciting version (the traditional version of the game) and a statement-eliciting version. In the statement-eliciting version, players produced statements about their own mystery person while the other player listened and eliminated all characters that did not exhibit a particular feature. In the question-eliciting version, players asked questions about the other player's mystery person by asking about specific features of this person. Note that in Catalan statements and yes-no questions have the same word order and they are only distinguished by intonation, rising for questions and falling for statements (unlike in English, for instance, where there is also subject/verb inversion).

All utterances and gestures were spontaneously produced as a result of the natural interaction between players. Crucially for our goals, participants spontaneously produced utterances that had target words in broad focus position and that were immediately followed by an intonational phrase boundary because they were produced at the end of the intonational phrase (see Table I for examples of a dialogue).

3. Procedure

While being paired up with another native speaker, all participants played the two versions of the game. The order was counter-balanced across pairs and both versions took place consecutively. During the game, participant A had to request information from participant B in order to find out the mystery person on B's board (question-eliciting version), or had to provide information to participant B so that participant B could guess the mystery person on A's board (statement-eliciting version). Players took turns asking questions or producing statements about the physical features of the "mystery persons." The winner was the player who first guessed the other's mystery person. No specific instructions were given to participants on the type of utterances they had to produce or on specific gestures they could use.

Participants sat facing each other across a table and in front of two laptop computers arranged so that they could not see each other's screen. Participants were audio-visually recorded using two Panasonic HD AVCCAMs at 50 frames per second. The cameras were placed on a tripod at a distance of approximately 1 m from the participants, each one facing a different member of the dyad. The cameras' height

TABLE I. Examples of a dialogue observed in the question-eliciting version of the game (dialogue 1) and in the statement-eliciting version of the game (dialogue 2). Words in bold are target prosodic words produced in broad focus position at the end of the prosodic phrase, and accented syllables are underlined.

Dialogue 1		Dialogue 2	
Player A:	És una **dona**?	Player A:	És un **home.**
	'Is it a woman'		'Is it a man'
Player B:	Sí.	Player B:	D'acord.
	'Yes'		'Ok'
Player A:	Porta **barret**?	Player A:	Porta **bigoti.**
	'Does she wear a hat?'		'He has got a moustache'

was adjusted to the participants' height in such a way that the recording area included the participants' upper body and head. Once the participants were seated, the experimenter explained the game and gave instructions about the procedure to be followed for each of the two variations. Altogether each version of the game lasted approximately 20 min.

4. Coding

All utterances about the physical properties of the mystery person were orthographically annotated and classified as being accompanied by a head movement or not. Whenever the annotator doubted on this classification, a conservative criterion was used, meaning that utterances were coded as not being accompanied by a head gesture. The types of head movements that were included in the analyses were *head nods* (following Poggi *et al.*, 2010, a head nod was any vertical head movement in which the head, after a slight tilt up, bends downward and then goes back to its starting point), *upward movements* (a head movement directed upward in the opposite direction from nodding), and *head tilts* (a head inclination or sideward movement) (see Wagner *et al.*, 2014, for a complete overview of the head gesture forms). All selected sentences had the form of verb + article + noun/adjective (the article being optional), as in the statement *Porta barret* "(S)he has a hat."

From the total amount of utterances produced by participants ($N = 492$), 111 utterances (22.6% of the total) were spontaneously accompanied by a head gesture. This proportion of gesture production per total amount of utterances is consistent with previous studies (e.g., Alexanderson *et al.*, 2013; Ferré, 2014). All head gestures co-occurred with the target word in the sentence (i.e., the content word featuring the physical property of the character, be it noun or adjective).

Table II displays the summary distribution of spontaneously produced utterances across participants, the amount of head gestures accompanying the target word, and the stress patterns of the target prosodic words. It illustrates that stress-initial (SW) target words were the most frequently produced, followed by monosyllabic words (S), and stress-medial words (WSW). The least frequent pattern was the stress-final (WS).

All utterances that were accompanied by a head gesture were further coded in terms of speech and gesture features. For gestures, we used ELAN annotation software, a tool that allows precise, frame-by-frame navigation through the video recording (Lausberg and Sloetjes, 2009). As Fig. 1 illustrates, head nods are characterized by a fall-rise movement that is generally preceded by an upward motion (see Ishi *et al.*, 2014, for a detailed description of the head nod shapes). For the gesture annotation we identified the following three points within the gesture movement: the *onset of the gesture* (the point where the head starts moving from its rest position), the *gesture apex* (the point where the bi-directional fall-rise head movement changes its direction), and the *end of the gesture* (the point where the gesture movement returns to its rest position).

For speech, we manually annotated the beginning and endpoints of the entire utterance, of the target prosodic word, and of the accented syllable within that target prosodic word (see Fig. 2). We used Praat (Boersma and Weenink, 2012) for speech coding, and Praat annotations were then imported into ELAN. The following criteria were used for speech segmentation: utterances were pause-bounded meaningful semantic units; target prosodic words were end-of-utterance content words (nouns or adjectives) forming a tone group bearing one word stress; and the accented syllable within the target prosodic word was the syllable within the prosodic word that carried the stress (and consequently the pitch accent of the entire utterance).

B. Results

For the analyses, the following dependent variables were taken into account: (1) the distance in time between the beginning of the gesture and the beginning of the prosodic word, (2) the distance in time between the end of the gesture

TABLE II. Summary of all the utterances produced, classified as a function of the participant, the presence of a speech-accompanying gesture, and the stress pattern of the target prosodic word.

	Target words without co-speech head gesture				Target words with co-speech head gesture				
Participant	WSW	WS	SW	S	WSW	WS	SW	S	Total
1	14	5	18	10	1	0	6	4	58
2	12	3	16	10	1	2	7	0	51
3	15	16	20	17	0	0	1	1	70
4	11	9	17	10	4	1	9	6	67
5	11	8	28	10	5	4	13	3	82
6	2	1	15	3	1	1	4	2	29
7	3	9	12	6	1	0	2	0	33
8	4	2	3	1	1	0	0	0	11
9	1	0	0	1	0	0	1	1	4
10	0	0	0	1	2	1	6	0	10
11	3	2	3	1	0	3	6	2	20
12	1	7	12	5	0	0	5	2	32
13	0	1	12	5	1	0	1	0	20
TOTAL	77	63	156	80	17	12	61	21	492

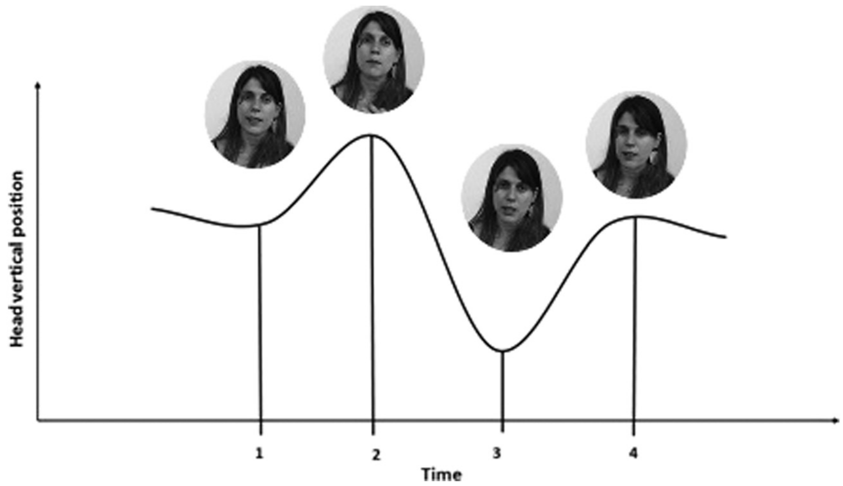

FIG. 1. Schematic representation of the relevant landmarks in a head nod gesture: the beginning of the gesture movement (1), the endpoint of the initial upward motion preceding the falling part of the movement (2), the gesture apex (3), and the end of the gesture (4). The preparation phase of the gesture corresponds to the temporal distance between points 1 and 2, the gesture stroke interval refers to the distance between 2 and 3, and the retraction phase interval is the distance between 3 and 4.

and the end of the prosodic word, and (3) the distance in time between the gesture apex and the end of the accented syllable. In all statistical analyses the fixed factor was the metrical pattern of the target prosodic word (4 levels: SW, WS, WSW, S), and the random factors were participant and item (simple random effects structure). Variables were assessed with linear mixed-effects models, using the *lmer* function within the lme4 package in R (Bates *et al.*, 2011). The models predicting the first two dependent variables will reveal what is the scope of the gesture movement, and whether it varies as a function of the position of the accented syllable and of the phrase boundary. The model predicting the third dependent variable will show if the gesture apex is produced within the temporal limits of the accented syllable, and whether the position of the intonational phrase boundary influences the precise location of the apex within this accented syllable.

Table III summarizes the results of the mixed-effects models. Results showed that the stress pattern of the prosodic word did not influence the distance between the gesture start and the start of the prosodic word or the distance between the gesture end and the end of the prosodic word. This means that, independently of the position of the prosodic prominence and of the upcoming phrase boundary, head movements started several milliseconds before the prosodic word started, and ended several milliseconds after the prosodic word ended (for descriptive values of all the analyses, see the Appendix). Instead, the stress patterns significantly impacted the temporal distance between the gesture apex and the end of the accented syllable, in that the stress-final patterns (S and WS) differed significantly from non-final stress patterns (SW and WSW). As Fig. 3 shows, the apex was aligned towards the middle of the accented syllable when there was non-accented material preceding the right-edge phrase boundary (SW and WSW), while it was much more retracted when the end of the accented syllable coincided with the presence of a right-edge phrase boundary (S and WS).

Three additional linear mixed-effects analyses with the same dependent variables and random factors were conducted, but now with sentence type as fixed factor (2 levels: question, statement). They revealed that the alignment patterns did not vary significantly as a function of this parameter (temporal distance between word start and gesture start: $\beta = 0.09$, $t = 1.33$; temporal distance between word end and gesture end: $\beta = 0.02$, $t = 0.14$; temporal distance between apex and end of accented syllable: $\beta = 0.07$, $t = 1.04$).

C. Discussion

In experiment 1 participants took part in two variants of the Guess Who game (one designed to elicit questions and the other to elicit statements), while being audio-visually recorded. Our aim was to see how speakers temporally

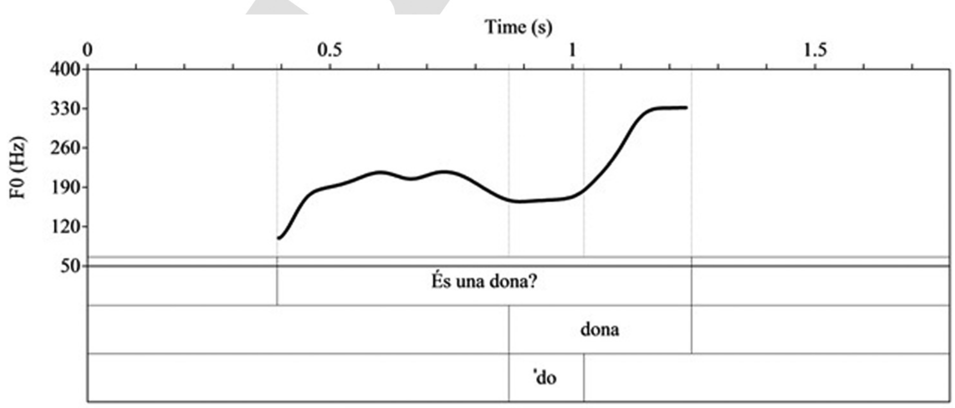

FIG. 2. Speech annotation of the utterances accompanied by a head gesture in Praat. First tier, temporal limits of the entire utterance. Second tier, temporal limits of the target prosodic word. Third tier, temporal limits of the accented syllable within that prosodic word.

TABLE III. Summary of the liner mixed-effects analyses for each dependent variable in experiment 1. Significant comparisons are in bold (we considered statistical significance to be $p \leq 0.05$).

	ß	SE	*t*
Gesture onset / word onset			
S vs WS	0.091	0.119	0.761
S vs SW	0.059	0.087	0.682
S vs WSW	0.099	0.113	0.881
WS vs SW	−0.031	0.104	−0.307
WS vs WSW	0.008	0.126	0.067
SW vs WSW	0.050	0.096	0.420
Gesture end / word end			
S vs WS	−0.039	0.183	−0.216
S vs SW	−0.194	0.133	−1.460
S vs WSW	−0.092	0.172	−0.535
WS vs SW	−0.154	0.157	−0.983
WS vs WSW	−0.052	0.192	−0.275
SW vs WSW	0.102	0.145	0.700
Gesture apex / end accented syllable			
S vs WS	−0.106	0.117	−0.905
S vs SW	0.257	0.085	**3.023**
S vs WSW	0.248	0.110	**2.245**
WS vs SW	0.363	0.101	**3.608**
WS vs WSW	0.354	0.123	**2.882**
SW vs WSW	−0.009	0.093	−0.102

aligned the head movements with speech while spontaneously interacting with an interlocutor. Specifically, we were interested in the influence of the prosodic heads (accented syllables) and phrase boundaries on the timing of head gestures.

The first main result was that speakers spontaneously produced head gestures together with the target prosodic

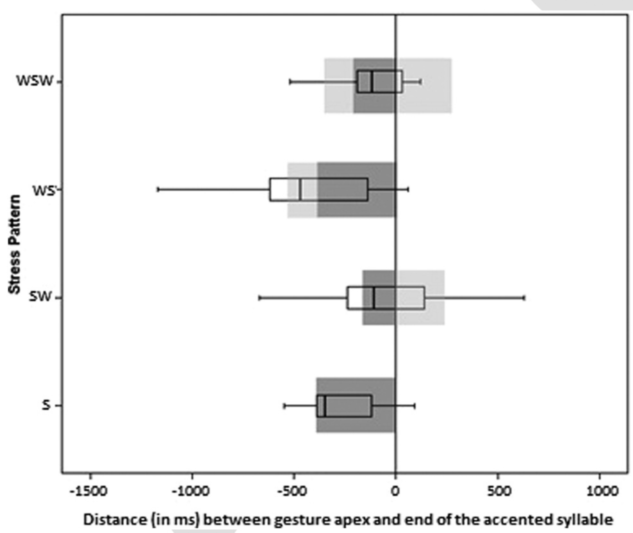

FIG. 3. Box plots displaying the temporal distance (in ms) between the gesture apex and the end of the accented syllable. The 0 represents the end of the accented syllable. Negative values show cases where the apex occurred before the end of the accented syllable. The dark grey shadow on top of box plots indicates the temporal limits of the accented syllable (means values) and the light grey shadows indicate the temporal limits of the non-accented syllables within the prosodic word (means values).

word. Participants were neither instructed regarding the type of sentences to be produced and were not explicitly told to gesture. Yet, all utterances included a phrase-final target word in broad focus position, and almost one fourth of the phrase-final target words were accompanied by a head gesture (head nod, head tilt, or upward movement). Despite the inter-individual variability in gestures production (also observed in Graf *et al.*, 2002; Ishi *et al.*, 2014; Swerts and Krahmer, 2010), the ratio of head gesture per utterance is similar to what previous studies have found when examining spontaneous interactions (Alexanderson *et al.*, 2013; Ferré, 2014) and indicates that the procedure was useful for the purposes of our study. Spontaneous data are valuable because they reveal the patterns of real-world interactions, but at the same time they complicate the examination of whether this variability is the result of different speaking styles or maybe of different pragmatic functions served by the head gesture (see experiment 2, and also the end of this section for a discussion of this issue).

The second main result was that the scope of the head gestures was the focused prosodic word. Irrespectively of the position of the prosodic prominence within the prosodic word, head gestures start close to the beginning of the corresponding prosodic word and they end after prosodic words are finished. This result contradicts those observed by Kim *et al.* (2014), who found that head movements occurred during the critical focused word in narrow-focus conditions but they occurred everywhere in broad-focus conditions. Yet, it goes in line with previous studies on gesture-speech alignment, which observed that the onset and offset of gesture movements are aligned with the onset and offsets of affiliated target words (e.g., Butterworth and Beattie, 1978; Kendon, 1980; Nobe, 2000; Roustan and Dohen, 2010; Schegloff, 1984).

The third main result, and in our view the most interesting one, refers to the temporal alignment of the gesture apex with the accented syllable. We found that the position of the head apex (the peak of gesture prominence) was influenced by the position of the accented syllable and of the upcoming phrase boundary. First, gesture apexes were produced within the temporal limits of the accented syllable (except for the WS case, in which the apex occurred during the pre-accented interval). Second, the exact anchoring point of the apex within the accented syllable depended on the position of the upcoming phrase boundary: the gesture apex was retracted if the prosodic word had the stress in phrase-final position (as in S and WS, possibly due to the prosodic pressure exerted by the upcoming prosodic boundary), and it was lagged if the prosodic word did not have the stress in phrase-final position (as in SW and WSW, where there is enough post-accentual material where the retraction of the head movement can be accommodated). The case of the phrase-final WS stress pattern is interesting because the apex is so retracted that it is produced out of the temporal limits of the accented syllable, suggesting that the position of the upcoming intonational phrase boundary has a stronger impact than the position of the accented syllable.

In sum, results from experiment 1 reveal that focused prosodic words determine the scope of head movements, that accented syllables seem to attract the peak of the gesture movement, and that phrase boundaries seem to determine the position of the peak within the accented syllable. The results of the WS patterns might also suggest that the effect phrase boundary might be stronger than that of the accented syllable. Thus, the prosodic structure of the utterance seems to have a strong impact on the timing of the apexes of speech-accompanying head gestures. This effect is consistent with previous results on the alignment of pitch peaks in rise-fall intonation contours (Prieto and Ortega-Llebaria, 2009), and of gesture peaks in manual pointing gestures (De Ruiter, 1998; Esteve-Gibert and Prieto, 2013).

However, some caveats in this experiment prevent us from drawing strong conclusions, mostly as a consequence of the spontaneous nature of the corpus. First, the spontaneous corpus yielded an unbalanced number of cases within each stress pattern condition. The results for the SW pattern, for instance, were based on a substantial number of cases, but the other patterns were three to five times less frequent. Second, although we controlled for sentence type (yes-no question versus statement), the spontaneous elicitation procedure did not allow us to finely control for the speakers' pragmatic intent. Previous studies have found that head nods can have different communicative functions: inclusivity, intensification, uncertainty, agreement, approval or emphasis (McClave, 2000; Poggi *et al.*, 2011; Poggi *et al.*, 2010). The emphatic function of head nods has also been observed in perception studies. It has been found that eyebrow movements and head nods help listeners to perceive prominent events in speech (House *et al.*, 2001; Krahmer and Swerts, 2007) and facilitate the recognition of prosodic contrastive focus (Dohen and Loevenbruck, 2004; Prieto *et al.*, 2015). It has been proposed that the temporal patterns of the gesture-speech integration can be influenced by semantic and pragmatic reasons (e.g., Bergmann *et al.*, 2011; Esteve-Gibert *et al.*, 2014). It could well be that the participants in our game responded with different degrees of commitment to the proposition and with different pragmatic intentions in mind. Maybe in experiment 1 the speaker's pragmatic intention had influenced the temporal alignment of the gesture-speech landmarks. Third, we do not know if the "attraction effect" of the accented syllable over the gesture apex is still maintained when there are larger distances between the accented syllable and the upcoming phrase boundary. It could be that this effect is reduced, maybe leading to gesture apexes that occur during the post-accented material. Experiment 2 was designed to remedy these concerns.

III. EXPERIMENT 2

The purpose of experiment 2 was to find additional support for the findings obtained in experiment 1. We designed a more controlled setting that would allow us to elicit head nod gestures with a co-referential meaning of confirmation, accompanying target words with specific stress patterns, and a balanced number of cases per stress pattern. Furthermore, an additional measure was taken into account in order to disentangle whether phrase boundaries have a stronger impact than accented syllables in determining the alignment of head gesture apexes with speech: the temporal distance between the beginning of the gesture stroke and the beginning of the accented syllable. This new measure will show us if the position of the prominent gesture interval (the gesture stroke) is determined by the position of the prosodic head (the accented syllable), by the upcoming phrase boundary, or by the entire prosodic word. Finally, in order to test whether the "attraction effect" of prosodic heads over gesture apexes is maintained when these heads are more distant to prosodic edges, a new stress pattern condition was included in the analyses (namely strong-weak-weak words, hereafter SWW).

A. Method

1. Participants

Eleven Catalan speakers (4 male, 7 female), between 22 and 54 years of age (mean age 30.5 years) participated in this experiment. All of them were students or staff at the Universitat Pompeu Fabra in Barcelona. They participated voluntarily and were not aware of the purpose of the experiment. None of them had participated in experiment 1.

2. Materials

Speakers were asked to participate in a Discourse Completion Task (DCT; Billmyer and Varghese, 2000; Blum-Kulka *et al.*, 1989) involving a set of 25 discourse contexts. A set of 25 cards was created, each containing a situation in which a hypothetical interlocutor is not sure whether a certain city (whose name appeared on the card) is the capital of a foreign country, a Spanish autonomous community, or a particular district in Catalonia. We chose to use names of world capital cities (and cities in Catalonia that would be well-known to all participants) as target words so that the situations described in the DCTs would be as close as possible to natural conversational situations.

Example (1) shows an example of a DCT. In this instance the target word is *Roma* "Rome," as indicated by the boldface.

(1) *Esteu jugant al Trivial i tu i en Joan sou part del mateix equip. Surt una fitxa que demana la capital d'Itàlia. En Joan en aquell moment dubta si la capital d'Itàlia és Roma i t'ho diu dubtant. Tu li dius que és cert, que és Roma, la capital d'Itàlia.*
Expected answer: *Sí, sí, la capital d'Itàlia és* ***Roma***.
"You and Joan are playing Trivial Pursuits and you are on the same team. The card you get asks you to name the capital of Italy. Joan is unsure and asks you whether it is Rome or not. You tell him that yes, it is Rome."
Expected answer: "Yes, yes, the capital of Italy is ***Rome***."

All of the discourse contexts used for the DCT task were designed to elicit a declarative sentence expressing confirmation. The target words had one of five different stress patterns, as described in Table IV. There were five target words for each pattern and they were expected to occur at the end of prosodic phrases. Each metrical pattern was chosen to represent a different position of prosodic prominence and prosodic edges, with stressed syllables in word

TABLE IV. The different stress patterns of the Catalan target words controlled for in experiment 2. In the examples column, stressed syllables are underlined.

Stress patterns of the target word	Position of the prosodic prominence	Examples
S	initial and final	Vic, Valls
WS	final	París, Dakar
SW	initial	Roma, Lima
SWW	initial	Mònaco, Washington
WSW	medial	Figueres, Caracas

initial, medial, or final position, and with unaccented syllables preceding, following, or surrounding the accented syllable.

3. Procedure

Participants were presented with one card at a time in random order, and were asked to read it carefully, to imagine themselves in the situation described in the discourse context, and, finally, to provide an appropriate verbal response. When participants provided a response that did not include the target word (e.g., *Sí, sí, és veritat* "Yes, yes, that's right"), the experimenter asked them to provide another response using the name of the capital city within the sentence. In order to elicit head nods as spontaneously as possible, participants were asked to produce spontaneous responses and were never prompted to gesture or produce utterances in an "expressive" manner.

Participants were audio-visually recorded using a Panasonic HD AVCCAM at 50 frames per second. The camcorder was placed on a tripod at a distance of approximately 1 m from the participant, and its height was adjusted to the participant's height in such a way that the recording area included the participant's upper body and head. The participants were recorded while standing up and were asked not to hold the DCT cards while providing a response. The entire procedure lasted approximately 15 min. A total of 275 trials (11 participants × 5 stress patterns × 5 items per pattern) were elicited.

4. Coding

We selected all utterances that were produced with a head nod gesture accompanying the target prosodic word, which occurred in focus position and was immediately followed by a prosodic boundary. The criterion for including head nods was the same as in experiment 1. From the total amount of trials ($N = 275$), 155 trials (56.4% of the total) were produced with a confirmation head nod gesture accompanying the target prosodic word. The remaining 120 trials were excluded from our analysis because speakers did not produce any head nod ($N = 48$), produced repetitive head nods associated with the adverb(s) *sí* "yes" and that continued during the entire utterance (called "hybrid" gestures in Yasinnik *et al.*, 2004) ($N = 39$), the target word was mispronounced ($N = 3$), or due to experimenter error ($N = 3$). We also excluded instances of head nods that co-occurred with the copular verb *és* "is" instead of with the target prosodic word ($N = 27$). Although these latter cases were pragmatically appropriate in the context of the task, they would have been included in the group of head nods accompanying monosyllabic S words and thus they would have unbalanced the number of trials per stress pattern.

Responses analyzed in this study had one of the following two structures: in 96.2% of the trials ($N = 149$) the target name was produced in the main clause at the end of the prosodic phrase (e.g., *Sí, sí, la capital de França és* París. "Yes, yes, the capital of France is Paris") and in 3.8% of the trials ($N = 6$) the target name appeared in a left-dislocated position, also at the end of the prosodic phrase (e.g., *Sí, sí, és* París*, la capital de França*. "Yes, yes, it is Paris, the capital of France").

All 155 valid trials were annotated in terms of speech and gesture. The speech annotation was the same as in experiment 1. The gesture annotation was very similar to experiment 1 except with the addition of an extra temporal landmark: the onset of the gesture stroke (point 2 in Fig. 1). As a result, four points within the head movement were identified in experiment 2: the onset of the gesture (the point at which the head starts moving from its rest position, the onset of the gesture stroke (the start of the falling part of the head movement), the gesture apex (the point in which directions change), and the end of the gesture (the point in which the gesture movement returns to its rest position).

B. Results

The following dependent variables were assessed using linear mixed-effects models (*lmer* function of the lme4 package in R, Bates *et al.*, 2011): (1) the start of the head movement with respect to the start of the target prosodic word, (2) the end of the head movement with respect to the end of that prosodic word, (3) the start of the gesture stroke with respect of the start of the accented syllable, and (4) the position of the gesture apex with respect of the end of the accented syllable. The fixed factor in all the analyses was the metrical pattern of the target prosodic word (five levels: S, SW, SWW, WS, and WSW), and random factors were participant and item (simple random effects structure).

Table V summarizes the results of the analyses and Fig. 4 illustrates these results in a visually succinct way. First, results revealed that the gesture started before the onset of the target word, and that the temporal distance between the two landmarks was the same across conditions. Only the stress-medial WSW pattern differed: compared to the other patterns, the gesture start was slightly closer to the word start (for descriptive values of all the analyses, see the Appendix). All target words in the elicited sentences were preceded by the copular verb *és* "is," hence gesture events that preceded the target prosodic word occurred during this preceding speech material.

Second, the temporal distance between the beginning of the gesture stroke and the beginning of the accented syllable varied significantly depending on whether there was preaccented material within the prosodic word, as it occurred closer to the beginning of the accented syllable in stress-initial words (S, SW, and SWW) and further from it in stress-final and stress-medial patterns. Figure 5 illustrates

TABLE V. Summary of the linear mixed-effects analyses for each dependent variable in experiment 2. Significant comparisons are in bold (we considered statistical significance to be $p \leq 0.05$).

	ß	SE	*t*
Gesture onset / word onset			
S vs SW	10.01	29.00	0.345
S vs SWW	−11.63	28.58	−0.407
S vs WS	−10.68	30.27	−0.353
S vs WSW	69.70	29.94	**2.328**
SW vs SWW	−21.64	27.84	−0.777
SW vs WS	−20.69	29.74	−0.696
SW vs WSW	59.69	29.24	**2.041**
SWW vs WS	0.94	29.28	0.032
SWW vs WSW	81.32	28.80	**2.823**
WS vs WSW	80.373	30.56	**2.630**
Gesture end / word end			
S vs SW	−19.852	29.768	−0.667
S vs SWW	−88.267	29.295	**−3.013**
S vs WS	−8.208	31.106	−0.264
S vs WSW	−87.092	30.696	**−2.837**
SW vs SWW	−68.42	28.50	**−2.400**
SW vs WS	11.64	30.66	0.380
SW vs WSW	−67.24	30.00	**−2.241**
SWW vs WS	80.069	30.163	**2.654**
SWW vs WSW	1.175	29.487	0.040
WS vs WSW	−78.884	31.442	**−2.509**
Stroke onset / onset accented syllable			
S vs SW	−1.517	21.614	−0.070
S vs SWW	1.326	21.287	0.062
S vs WS	−102.790	22.580	**−4.552**
S vs WSW	−47.114	22.306	**−2.112**
SW vs SWW	2.843	20.721	0.137
SW vs WS	−101.272	22.226	**−4.556**
SW vs WSW	−45.597	21.785	**−2.093**
SWW vs WS	−104.116	21.875	**−4.760**
SWW vs WSW	−48.440	21.440	**−2.259**
WS vs WSW	55.68	22.81	**2.440**
Gesture apex / end accented syllable			
S vs SW	280.63	20.87	**13.449**
S vs SWW	285.94	20.53	**13.925**
S vs WS	−10.15	21.80	−0.465
S vs WSW	235.15	21.52	**10.929**
SW vs SWW	5.309	19.978	0.266
SW vs WS	−290.779	21.493	**−13.529**
SW vs WSW	−45.485	21.029	**−2.163**
SWW vs WS	−296.088	21.145	**−14.003**
SWW vs WSW	−50.794	20.668	**−2.458**
WS vs WSW	245.29	22.04	**11.129**

that this distance varied as a function of whether the onset of the prosodic word coincided with the accented syllable or not, since speakers always aligned the gesture stroke some milliseconds before the onset of the prosodic word.

Third, regarding the temporal distance between the end of the gesture and the end of the prosodic word, we found that the gesture end was aligned significantly differently in the trisyllabic words (SWW and WSW) compared to the other patterns (S, WS, and SW): in trisyllabic words the gesture end occurred a little before the end of the prosodic word, while in the other patterns it occurred closer to it.

Finally, the position of the gesture apex with respect to the end of the accented syllable differed depending on whether there was unaccented material preceding the phrase boundary. Stress-final (S and WS) patterns differed from stress-initial (SW and SWW) and stress-medial WSW patterns. Figure 6 shows that the gesture apexes occurred during the temporal limits of the accented syllable, but that their precise alignment within that syllable varied depending on the presence of unaccented material preceding the phrase boundary. Thus, the gesture apex was largely retracted when the accented syllable occurred in phrase-final position (S and WS patterns), but was produced towards the middle of the accented syllable when there was post-accentual material preceding the right-edge phrase boundary (SW, SWW, and WSW patterns).

C. Discussion

Three main results can be observed from experiment 2. First, we could confirm that the scope of a confirmatory head nod gesture is the accompanying focused prosodic word, not the accented syllable. This is evidenced by the fact that speakers start head movements several milliseconds before the prosodic word and end them several milliseconds before the prosodic word is finished. Speakers maintain these patterns even if there are strong edge constraints within the prosodic word (i.e., the prosodic word being initiated or finished with an accented syllable, as in the S, WS, SW, and SWW items). Likewise, when speakers produce a gesture together with a prosodic word that is less constrained in its edges (as in the WSW condition), these general patterns are maintained although with minor variations: the gesture onset is slightly closer to the word onset and the end of the gesture is slightly more distant to the end of the word.

Second, we found that the position of the peak of prominence in the gesture (the gesture apex) is sensitive not only to the position of the accented syllable (which had been found in many previous studies; Fernández-Baena *et al.*, 2014; Graf *et al.*, 2002; Hadar *et al.*, 1983; Ishi *et al.*, 2014), but that it is also highly sensitive to the distance to the upcoming intonational phrase boundary. The position of the accented syllable within the prosodic word determined where the gesture apex will be produced (because gesture apexes tend to occur within its limits). But the specific position of the apex within the accented syllable depended on the upcoming prosodic phrase boundary, because the position of the gesture apex is adapted to the presence or absence of post-accentual material: the gesture apex occurred closer to the end of the accented syllable when there were one or more unaccented syllables before the upcoming prosodic boundary; instead, the apex was retracted if the upcoming prosodic boundary occurred immediately after the accented syllable.

Third, complementary evidence regarding the important role of the prosodic word as the domain of head nod movements comes from the timing of the start of the gesture stroke, which in our data is associated with the left-edge of the prosodic word (e.g., where the word starts) rather than with the accented syllable. In our data, speakers started the gesture stroke before the beginning of the prosodic word, and thus the gesture stroke was aligned further from the

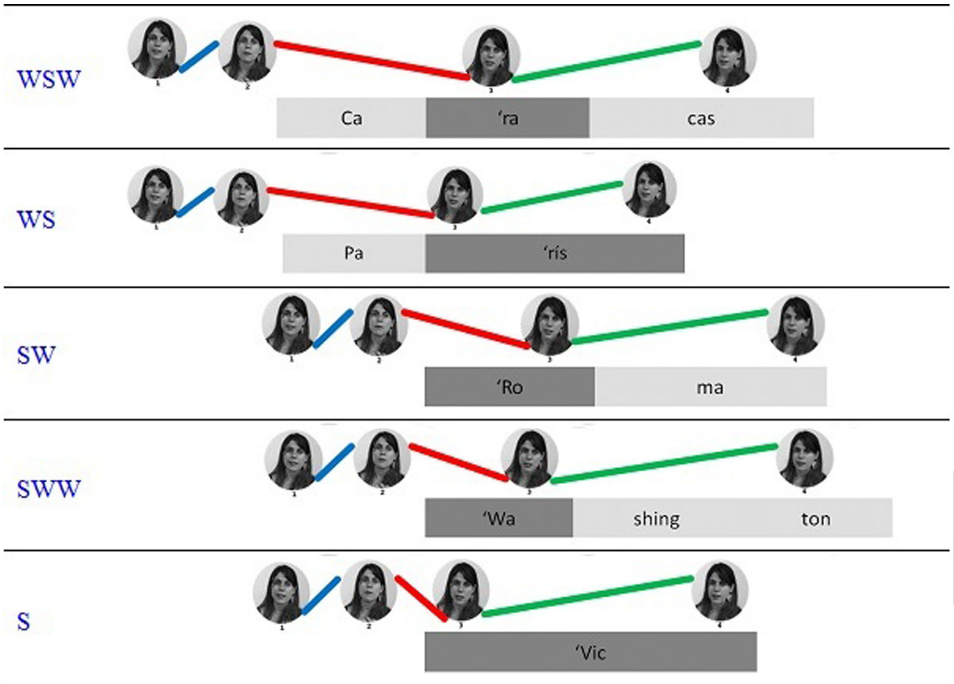

FIG. 4. (color online) Schematic representation of the alignment patterns of the head gesture and prosodic landmarks for each stress pattern. The dark grey cells represent the mean duration of the accented syllable within the prosodic word and the light grey cells the unaccented syllables. The lines connecting head images represent the gesture phases: the blue line from 1 to 2 is the preparation phase, the red line from 2 to 3 is the gesture stroke (the end of it being the gesture apex), and the green line from 3 to 4 is the retraction phase.

prosodic head in prosodic words with pre-accentual material (WS and WSW patterns), and closer to the start of the prosodic head when no pre-accentual material was available (e.g., S, SW, and SWW).

IV. GENERAL DISCUSSION AND CONCLUSION

The aim of this study was to investigate the effects of prosodic structure (i.e., the location of prosodic prominences and prosodic phrase boundaries) on the timing of head nod gestures. We designed two experiments, one that elicited spontaneous head gestures through a *Guess Who* game and another one that elicited semi-controlled head gestures in which we could better control for the speakers' communicative intent and the stress pattern of the target focused word. The results of experiment 1 showed that the scope of head movements is the whole prosodic word they accompany, and that the peak of the head movement (the gesture apex) occurs within the accented syllable of the prosodic word, its exact position depending on the presence or absence of an upcoming phrase boundary. A second experiment was required in order to refine and confirm these results, now (1) balancing the number of target prosodic words per stress pattern, (2) analysing a more complete set of stress patterns, (3) controlling for the speakers' communicative intent by eliciting confirmatory sentences, and (4) measuring also the

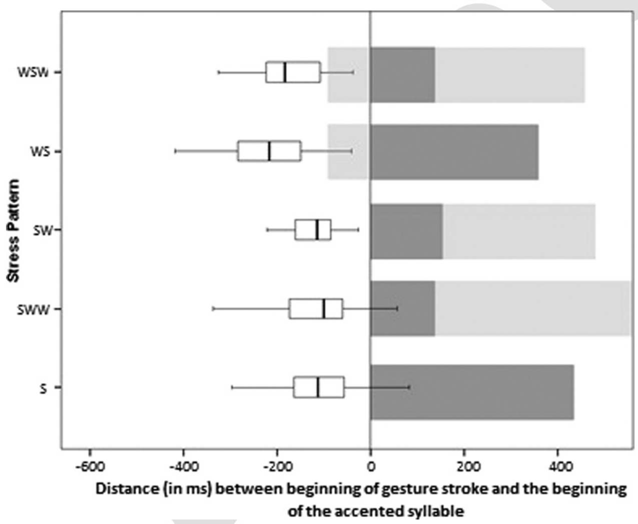

FIG. 5. Box plots displaying the temporal distance between the beginning of the gesture stroke and the beginning of the accented syllable. The 0 represents the beginning of the accented syllable, negative values showing cases where the gesture stroke started before the accented syllable and positive values the opposite. The dark grey boxes indicate the temporal limits of the accented syllable (mean values) and the light grey boxes indicate the temporal limits of the un-accented material within the prosodic word (mean values).

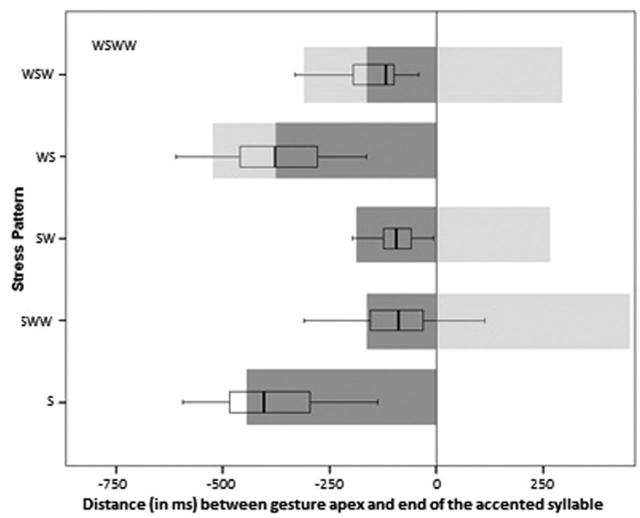

FIG. 6. Box plots displaying the temporal distance (in ms) between the gesture apex and the end of the accented syllable. The 0 represents the end of the accented syllable, negative values showing cases where the apex occurred before the end of the accented syllable and positive values the opposite. The dark grey boxes indicate the temporal limits of the accented syllable (mean values) and the light grey boxes indicate the temporal limits of the unaccented material within the prosodic word (mean values).

impact of the prosodic structure on the beginning of the prominent gesture interval, the gesture stroke.

Experiment 2 confirmed that the scope of the head movement is the accompanying focused prosodic word. Likewise, we found that the beginning of the prosodic word is the anchoring point for the start of the prominent interval of the gesture movement (the gesture stroke), hence moving it away from the accented syllable in prosodic words with pre-accented material. Crucially, we confirmed that the peak of the gesture movement, the apex, is timed as a function of the prosodic heads and edges: it occurs within the accented syllable independently of the metrical pattern of the target word, but its exact anchoring point within that syllable is retracted if there is an upcoming prosodic phrase boundary and lagged if there is post-accentual material before the prosodic phrase boundary occurs.

Previous research on the alignment of head gestures with speech had shown that accented syllables were the anchoring site for head apexes (Alexanderson *et al.*, 2013; Fernández-Baena *et al.*, 2014; Goldenberg *et al.*, 2014; Graf *et al.*, 2002; Hadar *et al.*, 1983; Ishi *et al.*, 2014). Yet, they also reported variability in this pattern. Our results suggest that an important source of variability is related to the position of prosodic edges, and specifically the distance between the accented syllable and the upcoming prosodic phrase boundary, a factor that none of these studies had controlled for. Previous research on pointing gestures had shown that the timing of pointing apexes resembles that of F0 movements (because pointing apexes align with F0 peaks, and these are retracted or lagged depending on the position of phrase boundaries) and of oral gestures (because manual gestures are lengthened at phrase boundaries) (Esteve-Gibert and Prieto, 2013; Krivokapić *et al.*, 2015; Krivokapić *et al.*, 2016; Rochet-Capellan *et al.*, 2008). Our results reveal that head movements are also affected by prosodic phrasing. This seems to be due to the fact that speakers plan the timing of their co-speech gestures by taking into account the prosodic features of the interval that will accommodate their associated gesture movements, and importantly the prosodic head and edge positions.

These results have direct implications for applied research. The temporal alignment of head gestures and speech is relevant for those researchers interested in designing virtual agents that interact in conversations as naturally as possible, the so-called "talking heads." Models of gesture-speech temporal integration should incorporate the effects of prosodic structure at several levels of speech planning. Research studying the semantic integration of gesture and speech has proposed that co-speech gestures refer to "lexical affiliates" (Schegloff, 1984). Here we propose that the temporal patterns of the gesture-speech alignment are explained by the impact of the different levels of the prosodic hierarchy on the planning and execution of the gesture movement.

Future studies should further investigate this entrainment between gesture and prosodic structure in speech. More work is needed to investigate how prosodic domains affect the temporal patterns in the realization of co-speech gestures. In our materials, for instance, we cannot disentangle whether the scope of the gesture movement is the lexical word or the prosodic word. Also, if prosodic structure strongly constrains the timing of head nod gestures (and co-speech gestures in general), speakers should have fine-grained perceptual expectations about gesture timing if a specific prosodic structure is predicted in the discourse. Finally, the influence of the semantic and pragmatic aspects of a gesture on its temporal implementation deserves further investigation, as recent studies examining spontaneously elicited gestures suggest that this influence can induce different types of gesture-speech temporal integration (e.g., Bergmann *et al.*, 2011; Esteve-Gibert *et al.*, 2014).

What seems to be beyond question is that there is tight temporal integration of gesture and speech, and that prosodic structure is one of the main aspects controlling this temporal coordination. Speakers use speech and gesture together to transmit their message, and discourse prominence is communicated at both the visual and acoustic levels by integrating the phases of gesture movements with the prosodic structure of oral messages.

ACKNOWLEDGMENTS

Thanks to Igor Jauk for the gesture coding in experiment 1, and Suleman Shahid and Constantijn Kaland for helping us with the experimental setting and recordings. This research has been funded by the Spanish MINECO (grant FFI2015-66533-P), and by the Generalitat de Catalunya to the Prosodic Studies Group (2014SGR-925), by the 2010 BE1 00207 travelling grant awarded to the second author of the study, and by the Labex BLRI (ANR-11-LABX-0036) grant awarded to the first author of the study.

APPENDIX

Descriptive results of all the analyses in experiments 1 and 2 are given in Table VI (all duration and distance measures are in milliseconds).

TABLE VI. Descriptive results of all the analyses in Experiments 1 and 2 (all duration and distance measures are in milliseconds).

	S	WS	WSW	SW	SWW[a]
Experiment 1					
Duration accented syllable	M = 434.5 (SD = 116.3)[b]	M = 420 (SD = 92.4)	M = 169.3 (SD = 35.3)	M = 164.2 (SD = 54.3)	—
Distance onset word / onset accented syllable	M = 0 (SD = 0)	M = −149.8 (SD = 40.2)	M = −124.1 (SD = 48.8)	M = 0 (SD = 0)	—

TABLE VI. (*Continued.*)

	S	WS	WSW	SW	SWW[a]
Distance offset accented syllable / offset word	M = 0 (SD = 0)	M = 0 (SD = 0)	M = −291.5 (SD = 94.9)	M = −251.2 (SD = 79.3)	—
Distance onset gesture / onset word	M = −335.3 (SD = 326)	M = −245.6 (SD = 339)	M = −236.4 (SD = 392)	M = −277.1 (SD = 346)	—
Distance offset gesture / offset word	M = 286.2 (SD = 599)	M = 249.4 (SD = 674)	M = 224.8 (SD = 492)	M = .098 (SD = 491)	—
Distance apex / offset accented syllable	M = −371.4 (SD = 352.7)	M = −482.8 (SD = 368.7)	M = −118.2 (SD = 309.1)	M = −116.9 (SD = 345.5)	—
Experiment 2					
Duration of the accented syllable	M = 431.2 (SD = 116.7)	M = 378.7 (SD = 92.3)	M = 149.9 (SD = 28.2)	M = 177.2 (SD = 46.9)	M = 149.9 (SD = 38.1)
Distance onset word / onset accented syllable	M = 0 (SD = 0)	M = −134.2 (SD = 37.8)	M = −132.7 (SD = 30.1)	M = 0 (SD = 0)	M = 0 (SD = 0)
Distance offset accented syllable / offset word	M = 0 (SD = 0)	M = 0 (SD = 0)	M = −288.3 (SD = 81.3)	M = −273.5 (SD = 76.5)	M = −382.1 (SD = 109.6)
Duration preparation phrase of the gesture	M = 164.3 (SD = 88.8)	M = 211.5 (SD = 102.9)	M = 177.5 (SD = 61.2)	M = 148.4 (SD = 72.7)	M = 179.3 (SD = 113.8)
Duration of the gesture stroke	M = 170.8 (SD = 53.6)	M = 221.1 (SD = 83.3)	M = 184.4 (SD = 58.1)	M = 204.9 (SD = 65.7)	M = 181.1 (SD = 52.2)
Duration retraction phase of the gesture	M = 247.9 (SD = 118.5)	M = 248.9 (SD = 109.2)	M = 227.4 (SD = 105.8)	M = 234.8 (SD = 111.1)	M = 266.1 (SD = 122.1)
Distance onset gesture / onset word	M = −290.2 (SD = 129.7)	M = −300.7 (SD = 132.7)	M = −221.8 (SD = 94.3)	M = −277.9 (SD = 114.4)	M = −301.4 (SD = 112.3)
Distance offset gesture / offset word	M = −138.3 (SD = 158.1)	M = −131.2 (SD = 138.4)	M = −203.3 (SD = 109.1)	M = −140.4 (SD = 89.1)	M = −206.8 (SD = 161.5)
Distance onset stroke / onset accented syllable	M = −125.9 (SD = 102.8)	M = −223.5 (SD = 109.5)	M = −177 (SD = 79.8)	M = −129.5 (SD = 72.5)	M = −122.1 (SD = 91)
Distance apex / offset accented syllable	M = −386.2 (SD = 115.2)	M = −381.1 (SD = 120.5)	M = −142.5 (SD = 77.6)	M = −101.7 (SD = 63.6)	M = −90.9 (SD = 91.3)

[a]This column is empty in experiment 1 because this stress pattern was not observed in Experiment 1.
[b]Mean, M; standard deviation, SD.

Ahmad, M. I., Tariq, H., Saeed, M., Shahid, S., and Krahmer, E. (**2011**). "Guess who? An interactive and entertaining game-like platform for investigating human emotions," in *Human Computer Interaction. Towards Mobile and Intelligent Interaction Environments*, Lecture Notes in Computer Science 6763, edited by J. A. Jacko (Springer, Berlin, Germany), Vol. 3, pp. 543–551.

Alexanderson, S., House, D., and Beskow, J. (**2013**). "Aspects of co-occurring syllables and head nods in spontaneous dialogue," in *Proceedings of 12th International Conference on Auditory-Visual Speech Processing (AVSP2013)*.

Ambrazaitis, G., Svensson Lundmark, M., and House, D. (**2015**). "Head movements, eyebrows, and phonological prosodic prominence levels in Stockholm Swedish news broadcasts," in *FAAVSP - The 1st Joint Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing*, Vienna, Austria, pp. 42–42.

Barkhuysen, P., Krahmer, E., and Swerts, M. (**2008**). "The interplay between the auditory and visual modality for end-of-utterance detection," J. Acoust. Soc. Am. **123**, 354–365.

Bates, D., Maechler, M., and Bolker, B. (**2011**). "lme4: Linear mixed-effects models using S4 classes [R package version 0.99375-39]," http://CRAN.R-project.org/package-lme4 (Last viewed January 27, 2017).

Bergmann, K., Aksu, V., and Kopp, S. (**2011**). "The relation of speech and gestures: Temporal synchrony follows semantic synchrony," in *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction*, pp. 1–6.

Bergmann, K., Kahl, S., and Kopp, S. (**2014**). "How is information distributed across speech and gesture? A cognitive modeling approach," Cognit. Processing **15**(1), S84–S87.

Billmyer, K., and Varghese, M. (**2000**). "Investigating instrument-based pragmatic variability: Effects of enhancing discourse completion tests," Appl. Linguist. **21**(4), 517–552.

Blum-Kulka, S., House, J., and Kasper, G. (**1989**). "Investigating cross- cultural pragmatics: An introductory overview," in *Cross-Cultural Pragmatics: Requests and Apologies*, edited by S. Blum-Kulka, J. House, and G. Kasper (Ablex, Norwood, NJ), pp. 1–34.

Boersma, P., and Weenink, D. (**2012**). "Praat: Doing phonetics by computer," http://www.praat.org/ (Last viewed July 25, 2016).

Butterworth, B., and Beattie, G. (**1978**). "Gesture and silence as indicators of planning in speech," in *Recent Advances in the Psychology of Language: Formal and Experimental Approaches*, edited by R. Campbell and G. T. Smith (Plenum Press, New York), pp. 347–360.

De Ruiter, J. P. (**1998**). "Gesture and speech production," doctoral dissertation, Katholieke Universiteit, Nijmegen, the Netherlands.

Dohen, M., and Loevenbruck, H. (**2004**). "Pre-focal rephrasing, focal enhancement and post-focal deaccentuation in French," in *Proceedings of the 8th International Conference on Spoken Language Processing*, pp. 2–5.

Esteve-Gibert, N., Pons, F., Bosch, L., and Prieto, P. (**2014**). "Are gesture and prosodic prominences always coordinated? Evidence from perception and production," in *Proceedings of the Speech Prosody Conference*, edited by N. Campbell, D. Gibbon, and D. Hirst, pp. 222–226.

Esteve-Gibert, N., and Prieto, P. (**2013**). "Prosodic structure shapes the temporal realization of intonation and manual gesture movements," J. Speech Language Hear. Res. **56**, 850–864.

Fernández-Baena, A., Montaño, R., Antonijoan, M., Roversi, A., Miralles, D., and Alías, F. (**2014**). "Gesture synthesis adapted to speech emphasis," Speech Commun. **57**, 331–350.

Ferré, G. (**2014**). "A multimodal approach to markedness in spoken French," Speech Commun. **57**, 268–282.

Goldenberg, D., Tiede, M., Honorof, D. N., and Mooshammer, C. (**2014**). "Temporal alignment between head gesture and prosodic prominence in naturally occurring conversation: An electromagnetic articulometry study," J. Acoust. Soc. Am. **135**, 2294.

Graf, H. P., Cosatto, E., Strom, V., and Huang, F. J. (**2002**). "Visual prosody: Facial movements accompanying speech," in *Proceedings of the 5th IEEE International Conference on Automatic Face Gesture Recognition*, pp. 396–401.

Guellaï, B., Langus, A., and Nespor, M. (**2014**). "Prosody in the hands of the speaker," Front. Psychol. **5**, 1–8.

Hadar, U., Steiner, T. J., Grant, E. C., and Rose, F. C. (**1983**). "Kinematics of head movements accompanying speech during conversation," Human Movement Sci. **2**(1–2), 35–46.

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

Ishi, C. T., Ishiguro, H., and Hagita, N. (**2014**). "Analysis of relationship between head motion events and speech in dialogue conversations," Speech Commun. **57**, 233–243.

Jannedy, S., and Mendoza-Denton, N. (**2005**). "Structuring Information through Gesture and Intonation," Interdisciplinary Stud. Inf. Struct. **3**, 199–244.

Kelly, S. D., Ozyürek, A., and Maris, E. (**2010**). "Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension," Psychol. Sci. **21**(2), 260–267.

Kendon, A. (**1980**). "Gesticulation and speech: Two aspects of the process of utterance," in *The Relationship of Verbal and Nonverbal Communication*, edited by M. R. Key (Mouton, the Hague, the Netherlands), pp. 207–227.

Kim, J., Cvejic, E., and Davis, C. (**2014**). "Tracking eyebrows and head gestures associated with spoken prosody," Speech Commun. **57**, 317–330.

Krahmer, E., and Swerts, M. (**2007**). "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception," J. Mem. Language **57**(3), 396–414.

Krivokapić, J. (**2014**). "Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes," Philos. Trans. R. Soc. London Ser. B Biol. Sci. **369**(1658), 20130397.

Krivokapić, J., Tiede, M. K., and Tyrone, M. E. (**2015**). "A kinematic analysis of prosodic structure in speech and manual gestures," in *Proceedings of the 18th International Congress of Phonetic Sciences*.

Krivokapić, J., Tiede, M. K., Tyrone, M. E., and Goldenberg, D. (**2016**). "Speech and manual gesture coordination in a pointing task," in *Proceedings of the 8th International Conference on Speech Prosody*, pp. 1240–1244.

Lausberg, H., and Sloetjes, H. (**2009**). "Coding gestural behavior with the NEUROGES-ELAN system," Behav. Res. Methods Instrum. Comput. **41**(3), 841–849.

Leonard, T., and Cummins, F. (**2011**). "The temporal relation between beat gestures and speech," Lang. Cognit. Processes **26**(10), 1457–1471.

Levelt, W. J. M., Richardson, G., and La Heij, W. (**1985**). "Pointing and voicing in deictic expressions," J. Mem. Language **24**, 133–164.

Loehr, D. P. (**2012**). "Temporal, structural, and pragmatic synchrony between intonation and gesture," Lab. Phonol. **3**, 71–89.

McClave, E. Z. (**2000**). "Linguistic functions of head movements in the context of speech," J. Pragmatics **32**, 855–878.

McNeill, D. (**1992**). *Hand and Mind: What Gestures Reveal About Thought* (University of Chicago Press, Chicago, IL).

Nobe, S. (**2000**). "Where to most spontaneous representational gestures actually occur with respect to speech?," in *Language and Gesture*, edited by D. McNeill (Cambridge University Press, Cambridge, UK), pp. 186–198.

Özyürek, A., Willems, R. M., Kita, S., and Hagoort, P. (**2007**). "On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials," J. Cognit. Neurosci. **19**(4), 605–616.

Poggi, I., D'Errico, F., and Vincze, L. (**2011**). "68 Nods. But not only of agreement," in *68 Zeichen Für Roland Posner. Ein Semiotisches Mosaik*. (*68 Signs for Roland Posner. A Semiotic Mosaic*) (Stauffenburg Verlag, Tübingen, Germany).

Poggi, I., D'Errico, F., Vincze, L., and Milazzo, V. (**2010**). "Types of nods. The polysemy of a social signal," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Malta.

Prieto, P., and Ortega-Llebaria, M. (**2009**). "Do complex pitch gestures induce syllable lengthening in Catalan and Spanish?," in *Phonetics and Phonology: Interactions and Interrelations*, edited by M. Vigário, S. Frota, and M. J. Freitas (John Benjamins, Philadelphia, PA), pp. 51–70.

Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E., and Blat, J. (**2015**). "Exploring the contribution of prosody and gesture to the perception of focus using an animated agent," J. Phonetics **49**, 41–54.

Rochet-Capellan, A., Laboissière, R., Galván, A., and Schwartz, J. (**2008**). "The speech focus position effect on jaw-finger coordination in a pointing task," J. Speech Language Hearing Res. **51**(6), 1507–1521.

Roustan, B., and Dohen, M. (**2010**). "Gesture and speech coordination: The influence of the relationship between manual gesture and speech," in *Proceedings of 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, Makuhari, Japan.

Rusiewicz, H. L., Shaiman, S., Iverson, J. M., and Szuminsky, N. (**2013**). "Effects of prosody and position on the timing of deictic gestures," J. Speech Language Hear. Res. **56**(2), 458–470.

Schegloff, E. A. (**1984**). "On some gestures' relation to talk," in *Structures of Social Action*, edited by J. M. Atkinson and J. Heritage (Cambridge University Press, Cambridge, UK), pp. 266–298.

Shattuck-Hufnagel, S., Ren, P. L., and Tauscher, E. (**2010**). "Are torso movements during speech timed with intonational phrases?," in *Proceedings of the Speech Prosody 2010*, Chicago, IL.

Swerts, M., and Krahmer, E. (**2010**). "Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions," J. Phonetics **38**, 197–206.

Treffner, P., Peter, M., and Kleidon, M. (**2008**). "Gestures and phases: The dynamics of speech-hand communication," Ecol. Psychol. **20**(1), 32–64.

Wagner, P., Malisz, Z., and Kopp, S. (**2014**). "Gesture and speech in interaction: An overview," Speech Commun. **57**, 209–232.

Yasinnik, Y., Renwick, M., and Shattuck-Hufnagel, S. (**2004**). "The timing of speech-accompanying gestures with respect to prosody," in *Proceedings From Sound to Sense: 50+ Years of Discoveries in Speech Communication* (MIT, Cambridge, MA), pp. C97–C102.