EMPIRICAL STUDY

Beat Gestures and Syntactic Parsing: An ERP Study

Emmanuel Biau,^a Lauren A. Fromont,^{b,c}
and Salvador Soto-Faraco^{d,e}

^aUniversity of Maastricht, ^bUniversity of Montreal, ^cCentre for Research on Brain, Language and Music, ^dUniversitat Pompeu Fabra, and ^eInstitució Catalana de Recerca i Estudis Avançats

We tested the prosodic hypothesis that the temporal alignment of a speaker's beat gestures in a sentence influences syntactic parsing by driving the listener's attention. Participants chose between two possible interpretations of relative-clause (RC) ambiguous sentences, while their electroencephalogram (EEG) was recorded. We manipulated the alignment of the beat within sentences where auditory prosody was removed. Behavioral performance showed no effect of beat placement on the sentences' interpretation, while event-related potentials (ERPs) revealed a positive shift of the signal in the windows corresponding to N100 and P200 components. Additionally, post hoc analyses of the ERPs time locked to the RC revealed a modulation of the P600 component as a function of gesture. These results suggest that beats modulate early processing of affiliate words in continuous speech and potentially have a global impact at the level of sentence-parsing components. We speculate that beats must be synergistic with auditory prosody to be fully consequential in behavior.

Keywords audiovisual speech; gestures; prosody; syntactic parsing; ERPs; P600

Introduction

Spoken communication in conversations is often multisensory, containing both verbal as well as nonverbal information in the form of acoustic and visual

This research was supported by the Ministerio de Economía y Competitividad (PSI2016-75558-P), AGAUR Generalitat de Catalunya (2014SGR856), and the European Research Council (StG-2010 263145). EB was supported by a postdoctoral fellowship from the European Union's Horizon 2020 research and innovation programme, under the Marie Skłodowska-Curie grant agreement No. 707727.

Correspondence concerning this article should be addressed to Emmanuel Biau, University of Maastricht, FPN/NP &PP, PO Box 616, 6200 MD, Maastricht, Netherlands. E-mail: emmanuel.biau@maastrichtuniversity.nl

signals, often conveyed by the speaker's gestures while speaking (see Biau & Soto-Faraco, 2013; McNeill, 1992). This study focuses on the listener's neural correlates expressing the interaction between the sight of the speaker's rhythmic hand gestures, commonly called "beat gestures" (McNeill, 1992), and the processing of the corresponding verbal utterance. Beats are rapid flicks of the hand that do not necessarily carry semantic information and are considered as a visual support to prosody (Balaguer, & Soto-Faraco, 2015; Biau, Morís Fernández, Holle, Avila, & Soto-Faraco, 2016; Biau, Torralba, Fuentemilla, de Diego Holle et al., 2012). Compared to other types of speech gestures, beats are by far the most frequent in conversations. Yet, their function on the listeners' end—if any—is still poorly understood. This study addresses the listener's neural correlates to beat gestures synchronized to words appearing in syntactically ambiguous sentences, in order to test their potential role as prosodic cues to sentence structure during speech processing.

First, we aimed to validate previous event-related potential (ERP) findings under more controlled conditions. These findings suggest that gestures may operate as attention cues to particular words in the utterance (Biau & Soto-Faraco, 2013). In addition, we addressed the potential role that these attention-grabbing beat gestures might play as prosodic markers with a function in syntactic parsing (Holle et al., 2012). The prosodic role of gestures at a sentence-processing level would be mediated by the aforementioned attention-cuing effect, which would play out at a sensory and/or perceptual stage of processing. In Biau and Soto-Faraco's previous study, ERPs were recorded while viewers watched audiovisual speech from a real-life political discourse, from a TV broadcast. The ERPs, time locked to the onsets of words pronounced with an accompanying beat gesture, revealed a significant early modulation within the window corresponding to the P200 component. The ERPs were more positive when words were accompanied by a beat gesture, as compared to the same words uttered without gesture. This result was in line with the assumption that gestures are integrated with the corresponding word at early stages (within the time window of the N100 and P200 ERP components), similar to other kinds of audiovisual modulations involving speech sounds and their corresponding lip movements (Brunellière, Sánchez-García, Ikumi, & Soto-Faraco, 2013; Pilling, 2009; van Wassenhove, Grant, & Poeppel, 2005). These results suggested an effect of the beat at phonological stages of processing, possibly reflecting the visual emphasis on the affiliate words (Krahmer & Swerts, 2007). Such modulations putatively occurring at a phonological level of processing are in line with the idea that temporal correspondence between beat gestures and pitch modulations of the voice mutually support a potential impact in prosody (Krahmer & Swerts,

2007; Treffner, Peter, & Kleidon, 2008). Although relevant for its ecological validity, the approach in Biau and Soto-Faraco's study did not allow for full control of differences between the amount of visual information in the gesture versus no-gesture conditions, other than the hand gesture itself. Furthermore, word pairs were extracted from different parts of the discourse, so that their different syntactic and phonological context was not controlled for, and their acoustic differences were only controlled for a posteriori, during data analysis. In this study, we controlled for both visual and auditory information to contrast the processing of acoustically identical words that could be accompanied, or not, by a beat gesture in carefully controlled sentences. If beat gestures effectively affect the early stages of processing of their affiliate words as suggested in previous studies (Hubbard, Wilson, Callan, & Dapretto, 2009; Marsteller & Burianová, 2014; McNeill, 1992), we expected to find a modulation during the time window corresponding to N100 and P200 in the ERPs time locked to word onsets, compared to the same words pronounced without a beat gesture.

Second, this study addressed whether beat gestures have an impact on speech comprehension by modulating syntactic parsing via their impact as attention cues. Based on the attention modulation account, we hypothesized that the (temporal) alignment of a beat gesture in the sentence would have an effect on syntactic interpretation by summoning the listener's attention at critical moments (Krahmer & Swerts, 2007; Kelly, Kravitz, & Hopkins, 2004; McNeill, 1992). We reasoned that, as beat gestures are normally aligned with acoustic prosodic markers in natural speech (Treffner et al., 2008), they could contribute to modulating syntactic parsing by boosting the perceptual saliency of their affiliate words. As mentioned above, beats likely affect processing of aligned auditory information, as reflected by early ERP modulations (Biau & Soto-Faraco, 2013), putatively by driving attention (e.g., Hillyard, Hink, Schwent, & Picton, 1973; Näätänen, 1982; Picton & Hillyard, 1974). Auditory prosody, conveyed by pitch accent, lengthening, or silent breaks, has already been shown to facilitate online spoken comprehension by cuing the correct parsing of sentences (Cutler & Norris, 1988; Gordon & Lowder, 2012; Lehisté, 1973; Quené & Port, 2005). For example, prosodic breaks together with a rising of the fundamental frequency (f_0) help listeners segment the signal into intonational phrases and facilitate decoding the syntactic structure (Clifton, Carslon, & Frazier, 2002; Frazier, Carslon, & Clifton, 2006; Fromont, Soto-Faraco, & Biau, 2017). Remarkably, given their temporal alignment with prosodic modulations in the speaker's voice (f_0), beats have been hypothesized to be the visual expression of speech prosody and impact the perceived saliency of targeted words, even in the absence of acoustic markers of accent (Krahmers

& Swerts, 2007; Leonard & Cummins, 2012; McNeill, 1992). Gestures are initiated prior to their affiliate words' onset, and their apex (i.e., the functional maximum extension point of the movement) consistently aligns with the f_0 peak of the stressed syllable (Holle et al., 2012; Wang & Chu, 2013). Indeed, Holle et al. (2012) showed that when beats emphasize the critical word in sentences with complex structures, the P600 component (sensitive to sentence analysis difficulty) decreases (Haupt, Schlesewsky, Roehm, Friederici, & Bornkessel-Schlesewsky, 2008; van de Meerendonk, Kolk, Vissers, & Chwilla, 2010). Consequently, we hypothesized that visual information from beats might modulate the syntactic parsing of ambiguous sentences depending on their placement by summoning the listener's attention to the affiliate auditory word.

Scope of the Study

We used relative-clause (RC) ambiguous sentences from Fromont et al. (2017) composed of two noun phrases (NP1 and NP2) and a final RC that could be attached to either NP1 (high attachment [HA]) or NP2 (low attachment [LA]; for a review, see Fernández, 2003), such as in the following example: “*Someone shot [the servant]_{NP1} of [the actress]_{NP2} [who was on the balcony]_{RC}.” The sentence in this famous example has two interpretations, as either the servant (HA) or the actress (LA) could be the person on the balcony. In the previous study, it was already established that the position of an acoustic prosodic break is sufficient to change the preferred interpretation of these syntactically ambiguous sentences, when presented auditorily. Here, we used audiovisual versions of these sentences, in clips where the speaker's hands could be seen. Following the original auditory study, these ambiguous sentences were presented at the end of short stories told by a speaker who would use gestures throughout. In the critical sentence, at the end of each clip, the speaker produced a beat gesture either aligned with the first or the second noun (at NP1 or NP2). We used the cross-splicing technique to create the video stimuli to ensure that the auditory track was exactly the same across the different gesture conditions and that visual information varied only by the temporal position of an otherwise identical beat gesture (or its absence, in an additional baseline condition). First, we hypothesized that beats influence the early processing stages of their affiliate (temporally aligned) words and would therefore express in amplitude modulation of the ERP within the time window of the N100 and P200 components (Pilling, 2009; van Wassenhove et al., 2005). It should be noted that our stimuli consist of running naturalistic speech, and therefore the N100–P200 complex that is often seen clearly for isolated stimuli (sounds or words) might not be distinguishable in our ERPs recorded from words embedded in sentences. Hence,*

we will only refer to modulations occurring within the time windows typical of the N100 and P200 components, but cannot directly refer to component modulation. Second, we hypothesized that if beats provide reliable prosodic cues by virtue of attentional cuing, they could drive the listener's focus to the affiliate word in a way similar to acoustic markers of prosody (such as pitch modulation or prosodic breaks). If this is true, then we expect gestures to influence sentence interpretation, like the aforementioned acoustic prosodic markers. This influence should express as a change in the listeners' choice probability for sentence interpretation. In order to test these two hypotheses, we used a two-alternative forced choice (2AFC) task, combined with electroencephalograph (EEG) recordings with evoked potentials measured from the onset of the nouns in the relevant sentence NPs (see details below).

Method

Participants

Twenty-one native Spanish speakers (11 females, mean age: 23 ± 4 years) volunteered after giving informed consent, in exchange for 10€/h. All participants were right-handed and had normal or corrected-to-normal vision and no hearing deficits. Three participants were excluded from the ERP analysis after more than 35% of their EEG epochs was filtered out with automatic artifact rejection. The protocol of the study was approved by the Clinical Research Ethical Committee of the Parc de Salut Mar (Comité Ético de Investigación Clínica), from the University Pompeu Fabra.

Stimuli

Audio Materials

One hundred six RC sentences containing attachment ambiguity such as (1) were created:

- (1) *La policía arrestó [al protegido]_{NP1} [del mafioso]_{NP2} que paseaba.*

The police arrested **the protégé** of **the mobster** who was walking.

In order to keep stimuli as ambiguous as possible in the absence of prosodic cues, the RCs inserted in the sentences were shorter than four syllables (based on de la Cruz-Pavía, 2010). All NPs contained between three and five syllables including the determiner to ensure rhythmically similar stimuli across the set. Each experimental sentence was preceded by a context fragment to enhance naturalness and introduce a prosodic rhythm. In order to control for lexical effects, frequency, phonological neighbors, and familiarity¹ of NP1 and NP2 nouns were measured using *EsPal* (Duchon, Perea, Sebastián-Gallés, Martí, &

Carreiras, 2013). Levene's test revealed that the sample was homogeneous in all three dimensions ($p > .05$). An analysis of variance (ANOVA) with NP as a between-item variable (two levels: NP1, NP2) returned no significant effect of familiarity, $F(1, 91) = .768, p = .383$, and a marginally significant effect of phonological neighbors, $F(1, 151) = 3.186, p = .076$. Finally, paired t tests revealed no significant difference in log(frequency) between the two lists, $t(60) = 1.505, p = .138$. All sentences, with their contexts, were pretested in order to verify their ambiguity in a sample of six volunteer participants in a pretest. The volunteers were presented with the sentences in written form and were asked to choose an interpretation. Six sentences were excluded because they elicited an attachment preference (low or high) of more than 70% on average. In addition, following Grillo and Costa (2014), nine sentences were excluded because they displayed pseudo-relative small clauses characteristics, which have a bias toward HA (for a complete list, see Fromont et al., 2017). Two versions of each selected sentence were audio recorded using a unidirectional microphone MK600, Sennheiser, and the Audacity software (v. 2.0.3; sampling 24kHz). For each sentence, a female native speaker of standard Castilian Spanish was asked to read versions (2) and (3) in a natural fashion (“#” indicates a prosodic break).

(2) *La policía arrestó [al protegido]_{NP1} # [del mafioso]_{NP2} que paseaba.*

(3) *La policía arrestó [al protegido]_{NP1} [del mafioso]_{NP2} # que paseaba.*

Using Praat (Boersma & Weenink, 2015), the sentences were examined acoustically and visually (viewing the spectrograms) to make sure they presented homogeneous intonation. The two versions of sentences were then cross-spliced at the offset of the preposition (*‘del’*) to create one single version of the sentence, without prosodic break. For all soundtracks, we normalized the amplitude peaks to maximum, leading all average amplitudes in the files to be almost equal. In doing so, we equalized and normalized auditory material among sentences. The resulting sentences were judged to sound natural by the authors as well as three native Spanish speakers with phonetic training.

Video Editing

To create the video, a female actor (author LF) was video recorded while mimicking speaking over each auditory sentence (previously recorded by a different speaker). Note that the use of a different speaker and actor is only anecdotal because the materials had to be created by aligning the speaker's (gesture) videos with the corresponding auditory sentences recorded in another instance, for control reasons: In particular, we thought it was important to use

"[Context story]...The police arrested the protégé of the mobster who was walking."

Figure 1 Experimental procedure. (a) For each trial, participants attended to an audiovisual clip in which the speaker told a short story ending with a final ambiguous sentence. Depending on the condition, a beat gesture accompanied either the first noun (NP1, "protégé") or the second noun (NP2, "mobster"), or none. The video was followed by the two-alternative forced choice question to determine the final sentence interpretation (i.e., Who was walking? The "protégé" or the "mobster"). (b) The centro-parietal electrodes used for the event-related potential analysis (C3, C1, CP5, CP3, CP1, P3, P1, Cz, CPz, Pz, P2, P4, CP2, CP4, CP6, C2, and C4). [Color figure can be viewed at wileyonlinelibrary.com]

the same auditory soundtrack (obtained from a gesture-free pronunciation of the sentence) in all gesture conditions so no acoustic variables could explain differences in ERPs/behavior. For each video, the actor listened to the auditory track several times and read simultaneously its written transcription on a screen. Videos were recorded once she felt comfortable with the story and practiced enough to gesture naturally along with the speech's rhythm. The actor was instructed to gesture freely during the context fragment to improve the ecological aspect of the speaker's movements and avoid drawing the subjects' attention to the critical gesture in the final experimental sentence of each stimulus. For each trial, two videos were recorded: in the first one, the actor made a beat gesture aligned with NP2 of the experimental sentence (this video was later manipulated to create the condition where the gesture aligns with NP1). The actor always began the critical sentence of each stimulus with her hands in a standard position (i.e., placed at predefined markers on the table hidden from the viewer; see Figure 1) and went back to that standard position at the end of the final sentence. This allowed for manipulation of the number of frames between the onset of the last sentence and the onset of the gesture while maintaining the natural flow of the audiovisual clip when creating the other

gesture condition (i.e., gesture aligned with NP1). In the second version of each sentence, the actor did not execute any gesture during the final sentence.

From the two videos recorded for each sentence, we created three audio-visual conditions. The NP2 condition, in which the gesture is aligned with the second noun of the final sentence, was created by aligning the video with the corresponding audio track using Adobe Premiere Pro CS3. The NP1 condition, in which the gesture aligned with the first noun in the final sentence, was created from the NP2 condition by removing video frames when the speaker had her hands in the standard position at the onset of the final sentence, until the same gesture aligned temporally with NP1. Importantly, in both conditions, we aligned the beat's apex (i.e., the maximum extension point of the gesture) with the pitch (fundamental frequency) peak of the stressed syllable of the corresponding noun (measured in *Praat*; Boersma & Weenink, 2015). The baseline condition in which the speaker did not gesture during the final sentence was created by cross-splicing the videos of the NP1 and NP2 conditions, between the context and the experimental sentences. In doing so, we ensured that the visual information of the context was exactly the same across the three conditions for each story with the exception of a single beat gesture that was aligned on NP1 or NP2 in either condition. Because the actor's position with her hands at rest was kept constant, the cross-splicing point could be smoothed using a fading effect. The cross-splicing point always occurred between the context and the experimental sentences. After editing, the video clips were exported using the following parameters: video resolution 960×720 pixels, 25 fps, compressor Indeo video 5.10, AVI format; audio sample rate 48 kHz, 16 bits, Stereo. In all the AV clips, the face/head of the speaker was occluded from the viewer's sight in order to block visual information from the face/head, such as lip movements or head nods (see Figure 1).

Procedure

Participants sat on a comfortable chair in a sound-attenuated booth, about 60 centimeters from a monitor. Each trial started with a central white fixation cross displayed on a black background. The cross turned red and disappeared when the audiovisual stimulus started (context + final sentence). After the video ended, participants were prompted to choose between two interpretations of the last sentence of the clip with no time pressure (i.e., between NP1 and NP2). In order to ensure that participants attended to the whole speech content, they were also presented with a 2AFC comprehension question about the context sentence at the very end of the trial in 20% of the trials. We measured the reaction times and attachment preference rates.

EEG Recording and Preprocessing

Electrophysiological data were recorded at a rate of 500 Hz from 59 active electrodes (ActiCap, Brain Vision Recorder, Brain Products) whose impedance was kept below 10 k Ω , placed according to the 10–20 convention. Extra electrodes were located on the left/right mastoids and below and at the outer canthus of the right eye. An additional electrode placed at the tip of the participant's nose was used as a reference during recording. The ground electrode was located at the AFz location. Preprocessing was performed using BrainAnalyzer software (Brain Products). The data were rereferenced offline to the average of the mastoids. EEG data were filtered with a Butterworth (0.5Hz high-pass, 70Hz low-pass) and a notch filter (50Hz). Eye blink artifacts were corrected using the procedure of Gratton and Coles (1989). The remaining artifacts were removed applying automatic inspection on raw EEG data (amplitude change threshold at $\pm 70 \mu\text{V}$ within 200 milliseconds). When more than 35% of the epochs signal after the segmentation relative to triggers was marked as contaminated after the automatic inspection (12 epochs out of 33), the participant's data were removed from further ERP analysis. The data set was segmented into 600-millisecond epochs (from -100 milliseconds before, respectively, the NP1 and NP2 onsets to 500 milliseconds after the onsets). Baseline correction was performed in reference to the 100-millisecond window of prestimulus activity. In each condition, the grand average was obtained by averaging individual average waves. Based on our previous gesture studies (Biau et al., 2015; Biau & Soto-Faraco, 2013), we focused on the centro-parietal electrodes C3, C1, CP5, CP3, CP1, P3, P1, Cz, CPz, Pz, P2, P4, CP2, CP4, CP6, C2, and C4 for the ERP analysis (see Figure 1). We also placed triggers to measure ERPs time locked to the onset of the RC across conditions.

ERP Analysis

We ran separate analyses on the ERPs of each of the nouns corresponding to NP1 and NP2 of each sentence. For each word-evoked potential, two time windows were defined by hypothesis, before visual inspection of the signal, time locked to word onsets, regardless of condition. These windows were based on previous audiovisual integration studies looking at visual modulations of the auditory evoked potentials N100 and P200. We delimited a first time window from 60 to 120 milliseconds after word onset to inspect modulations around the time of the N100 component and a second time window from 170 to 240 milliseconds to capture modulations around the time when the P200 component usually occurs (Biau & Soto-Faraco, 2013; Pilling, 2009; Näätänen, 2001; Stekelenburg & Vroomen, 2007; van Wassenhove et al., 2005). For each time window (N100

and P200), mean ERP amplitudes in the three gesture conditions (gesture on NP1, gesture on NP2, and no-gesture baseline) for the electrodes of interest were extracted separately for each participant. Mean ERP amplitudes were then submitted to two-way (within-subjects) ANOVAs with the factors gesture condition (three levels: gesture on NP1, gesture on NP2, and no-gesture baseline) and electrode (seventeen levels: C3, C1, CP5, CP3, CP1, P3, P1, Cz, CPz, Pz, P2, P4, CP2, CP4, CP6, C2, and C4). The factors electrode and interaction electrode $\times$ gesture condition are not reported as we focus the analysis on the main effect of gesture on our cluster of electrodes based on previous literature, but not on the scalp distribution differences. Greenhouse-Geisser correction was applied to control for sphericity violations when appropriate. When the factor gesture condition was significant, a post hoc analysis using Bonferroni correction for multiple comparisons was applied to determine the pattern of the effect. For both time windows of interest (60–120 and 170–240 milliseconds), we performed peak detection on the average ERPs in the gesture conditions (gesture on NP1 and gesture on NP2) and reported the scalp distributions of the effects at peak timing in the three conditions. For the RC-evoked signal, we analyzed a time window of interest of 500–900 milliseconds after the onset of the RC. We performed the same analyses on the same electrode set as described above.

Results

Behavioral Results

Reaction Times (RT)

Participants were not under time pressure or given a time limit to respond so RTs are shown here for completeness. The analyses of RTs did not reveal any difference across conditions (NP1: 3053 ± 1287 milliseconds; NP2: 3045 ± 1231 milliseconds; Baseline: 3099 ± 1241 milliseconds). A one-way ANOVA with the factor gesture condition (three levels: gesture on NP1, gesture on NP2, and no-gesture baseline) did not show any significant effect, $F(2, 40) = 0.175$; $p = .84$.

Attachment Preference

Behavioral responses were classified in two categories: HA when participants attached the RC to NP1 and LA when the RC was attached to NP2. Figure 2 shows the modulations of HA preference across conditions. A one-way ANOVA on HA preference with the factor gesture condition (gesture on NP1, gesture on NP2, and no-gesture baseline) did not reveal any significant effect, $F(2, 40) = 0.967$; $p = .389$.

Figure 2 Modulation of high attachment preference depending on the gesture condition (gesture on NP1, gesture on NP2, and no-gesture baseline) in rates $\pm$ standard deviation. [Color figure can be viewed at wileyonlinelibrary.com]

ERPs

ERPs Measured to Noun NP1²

The ANOVAs revealed significant effects of gesture condition in both time windows: 60–120-millisecond window, $F(2, 34) = 6.45$, $p < .005$; 170–240-millisecond window, $F(2, 34) = 7.40$, $p < .005$. In both time windows, Bonferroni-corrected post hoc analyses showed that the amplitude of the ERP evoked by the NP1 word was significantly more positive when the gesture accompanied the NP1 word (gesture on NP1 condition), compared to when that same word was pronounced without gesture (no-gesture baseline and gesture on NP2; see Figure 3). Peak detection revealed a peak at 103 milliseconds in the 60–120-millisecond window and at 216 milliseconds in the 170–240-millisecond window, in the gesture on NP1 condition (Figure 3). The mean amplitudes in the two time windows of interest in the three conditions are also summarized in a bar graph (Figure S1 in the Supporting Information online). Additionally, we tested for the laterality of the effect of gesture condition by grouping electrodes in two regions of interest: left centro-parietal (C3, C1, CP5, CP3, CP1, P3, and P1) and right centro-parietal (P2, P4, CP2, CP4, CP6, C2, and C4). The two-way ANOVA with the factors gesture condition (three levels: gesture on NP1, gesture on NP2, and no-gesture baseline), laterality (left and right), and electrode (seven levels: C_x , C_y , CP_x , CP_y , CP_z , P_x , and P_y) did not reveal any significant effect of laterality or interaction gesture condition $\times$ laterality in the 60–120-millisecond or the 170–240-millisecond windows.

Figure 3 Top panel: Event-related potentials time locked to the onsets of (a) noun NP1, (b) noun NP2 at Cz site (in the sentence example, “*The police arrested the protégé of the mobster who was walking*”). The black line represents the signal when the gesture was aligned with NP1 (gesture on NP1 condition), the red line represents the signal when the gesture was aligned with NP2 (gesture on NP2 condition) and the blue line represents the signal when the final sentence was pronounced with no gesture (no-gesture baseline condition). Bottom panel: Scalp distributions at the peaks in the 60–120 millisecond and 170–240 millisecond time windows for (c) noun NP1 and (d) noun NP2. Peaks were detected in the two gesture conditions (gesture on NP1 and gesture on NP2) and we also report the scalp distribution in the other conditions at the equivalent time points. [Color figure can be viewed at wileyonlinelibrary.com]

ERPs Measured to NP2 Noun

The ANOVAs revealed significant effects of gesture condition in both time windows: 60–120-millisecond window, $F(2, 34) = 9.46$, $p < .001$, and 170–240-millisecond window, $F(2, 34) = 6.70$, $p < .005$. Bonferroni-corrected post hoc analyses showed that the ERP amplitude was significantly more positive when the gesture accompanied the NP2 word (gesture on NP2 condition),

Figure 4 Event-related potentials time locked to the onsets of the relative clause (in the sentence example “*The police arrested the protégé of the mobster [who was walking]_{RC}*”), at Cz site. The black line represents the signal when the gesture was aligned with NP1 (gesture on NP1 condition), the red line represents the signal when the gesture was aligned with NP2 (gesture on NP2 condition), and the blue line represents the signal when the final sentence was pronounced with no gesture (no-gesture baseline condition). [Color figure can be viewed at wileyonlinelibrary.com]

compared to when it was pronounced without gesture (no-gesture baseline and gesture on NP1 conditions; see Figure 3). Peak detection revealed a peak at 77 milliseconds in the 60–120 window and 178 milliseconds in the 170–240 window in the gesture on NP2 condition (Figure 3). The mean amplitudes in the two time windows of interest in the three conditions are also summarized in a bar graph (Figure S1 in the Supporting Information online). Again, no significant effect of laterality or interaction gesture condition $\times$ laterality was found in either the 60–120-millisecond or 170–240-millisecond windows.

ERPs Measured to RC

In addition, we executed a post hoc ERP analysis centered on a late time window locked to the RC across the three gesture conditions (see Figure 4). This analysis was not initially planned by hypothesis and aimed to explore whether the gestures may have had some impact on the P600 component. Even though our paradigm was not optimized to measure the P600 effect, as we did not manipulate sentence grammaticality, and although our behavioral results did not reveal any effect of gesture placement on the interpretation of ambiguous sentences, exploring late ERP responses to the RC may be of some

interest in this context given their well-known link to sentence parsing. The P600 component is commonly referred to as a measure of syntactic anomaly (Osterhout & Holcomb, 1992) in sentence processing and reanalysis (Friederici, 2002; including RC in Spanish, see Carreiras, Salillas, & Barber, 2004). In particular, a previous study found that beat gestures reduced the P600 amplitude of complex sentences (Holle et al., 2012).

The ANOVAs on the late ERP time locked to the RC revealed a significant effect of gesture condition on mean amplitudes in the time window of interest, 500–900 milliseconds, $F(2, 34) = 8.757$; $p = .001$. Bonferroni-corrected analyses showed a significant positive shift in the mean amplitude in the time period of interest in the no-gesture baseline condition, as compared to both NP1 and NP2 gesture conditions. Additionally, the analyses did not reveal any significant difference between the two gesture conditions (gesture on NP1 vs. gesture on NP2). The mean amplitudes in the two time windows of interest in the three conditions are also summarized in a bar graph (Figure S2 in the Supporting Information online).

In summary, the results showed that behavioral performance (choice probability) regarding the interpretation of ambiguous RC sentences was not affected by beat placement. In fact, the preference for the HA measure reflected that listeners' interpretations in either gesture condition was equivalent to their interpretation in the no-gesture baseline, which was fairly balanced, close to 50% (as intended by materials selection). In contrast, gestures exerted a strong effect in the early latency ERPs time locked to the onsets of their affiliate nouns in both NP1 and NP2 conditions. In both cases, the EEG signal time locked to the word (hence reflecting its processing) was modulated by the accompanying beat gesture with a significant positive shift in amplitude. More precisely, these robust effects were found in two time windows corresponding to the N100/P200 ERPs components. Finally, an additional post hoc analysis on the late latency RC-evoked signal (reflecting sentence processing stages), we found that the presence of a beat gesture, independently from its placement on NP1 or NP2, elicited a decrease in amplitude in the 500–900-millisecond time period. This reduction of the positive shift in the P600 window might reflect an ease of sentence processing when the sentences included gestures, as compared to the no-gesture baseline condition.

Discussion

The present study has addressed the neural correlates expressing the integration between the sight of beat gestures and the processing of the corresponding affiliate words, during processing of continuous audiovisual speech. The results

showed that the sight of beat gestures modulated the ERP responses to the corresponding word at early stages, corresponding to the time windows of the typical N100 and P200 auditory components. According to the significant time windows, one can contend that the gesture might have had an impact at different stages in audiovisual speech processing (Bart, Stekelenburg, & Vroomen, 2014; Brunellière et al., 2013; Pilling, 2009; van Wassenhove et al., 2005). Previous studies have related acoustic processing to the N100 and phonological processing to the P200 auditory components (e.g., Brunellière & Soto-Faraco, 2015; Brunellière et al., 2013; Brunellière & Soto-Faraco, 2013; Kong et al., 2010; Obleser, Scott, & Eulitz, 2006). Hence, this N1/P2 modulation supports the hypothesis that gestures may act as attention cues at early stages of speech processing and confirm previous results obtained with real-life stimuli (Biau & Soto-Faraco, 2013), using more controlled presentation conditions. Second, we measured the behavioral consequences of the alignment of beat gestures with critical words in sentences. We hypothesized that, by virtue of their attention effects, beats could have an impact on the interpretation of syntactically ambiguous sentences, similar to acoustic prosodic cues (Fromont et al., 2017). However, according to the behavioral results, choice probability (percentage of HA) did not reveal any modulation in sentence interpretations as a function of the position of the beat. Instead, the present behavioral results suggest that, at least in the absence of acoustic prosodic cues (i.e., such as pitch accent and breaks), listeners were not induced to prefer one interpretation or the other compared to the baseline (as reflected by HA choices around 50% in the three conditions).

Regarding the ERP results, both visual and auditory information were identical across conditions and varied only by the placement of a beat gesture in the final sentence of interest in each short story. Although we found no behavioral effect, the modulation in the word-evoked ERPs at the N100 and P200 time windows supports the account of the attentional effect of beats, which have been hypothesized to attract the listener's attention toward relevant information (Kelly et al., 2004; McNeill, 1992). Beats are often considered as highlighters, and listeners may rely on them to anticipate important words, owing to their predictive temporal relationship. Beats may modulate the visual context in a nonrandom manner, preparing the system for upcoming auditory inputs and having an effect on how they are processed. In line with this assumption, we previously showed that beat gestures induced the synchronization in the listeners' EEG in the theta band at (and even before) the onset of the affiliate word (Biau & Soto-Faraco, 2015; Biau et al., 2015), suggesting an anticipatory effect of gestures on the processing of the sensory input at relevant moments

in the auditory signal (Astheimer & Sanders, 2009). Cross-modal anticipation in audiovisual speech has been reported using several approaches, including ERPs (Brunellière & Soto-Faraco, 2015; Brunellière & Soto-Faraco, 2013; van Wassenhove et al., 2005) and behavior (Sánchez-García, Alsius, Enns, & Soto-Faraco, 2011), and seems to be one of the ways in which audiovisual integration confers a benefit in speech processing. Furthermore, the N100/P200 ERP components typically modulated in cross-modal prediction (anticipation) have been related to attentional modulations in seminal studies (Hillyard et al., 1973; Näätänen, 1982; Picton & Hillyard, 1974), but also more recently in auditory speech segmentation where their modulation was greater at relevant word onsets (e.g., Astheimer & Sanders, 2009). Although the precise peaks of the components are difficult to find when using words embedded in running speech, the modulations of the auditory evoked potential observed here occurred right at the time windows corresponding to these components (as guided by our *a priori* hypothesis).

From our viewpoint, there are at least two possible interpretations of the influence of beats on the neural correlates of the corresponding word. First, the effect of beats may simply reflect the extra visual input from the sight of hand gestures, as compared to conditions where the word was not accompanied by the gesture. One could claim, indeed, that this effect is unrelated to speech processing, as previous studies of visual perception of biological motion have reported early modulations in the ERP (Hirai, Fukushima, & Hiraki, 2003; Krakowski et al., 2011). For example, a study comparing biological to scrambled motion perception found a negative shift at latencies 200 and 240 milliseconds poststimulus onset, related to the higher processing of motion stimuli in the biological motion condition (Hirai et al., 2003). More recently, a study also showed that the percentage of biological motion contained in point-light animations modulated the amplitude of the N100 component, with the largest modulation corresponding to the full biological motion condition (Jahshan, Wynn, Mathis, & Green, 2015). However, despite being temporally compatible with our study's results, these findings do not fully explain the ERP modulations reported in our study. For instance, the modulations we observed in the N100 and P200 time windows consist of a positive shift, whereas biological motion perception often produces a negative shift on N100, compared to control conditions. Regarding scalp distribution, Jahshan et al. reported a posterior distribution of the N100 component when participants perceived full biological motion. Here, the scalp distributions in the gesture conditions (NP1 and NP2) in the N100 time window showed a widespread positive effect from centro-parietal to occipital sites, which can be in line with Jahshan et al. (2015).

However, even if biological motion correlates partially overlap with the ERP modulations seen here (although we defined our cluster of interest based on previous hypothesis and did not look at further electrodes), there are indications that beat–speech interactions are supported by tightly coupled timing (must occur at critical moments) and are not solely processed as noncommunicative body movements. Some evidence supports that speakers are experts at placing beat timing with their speech, and listeners seem to be highly sensitive at picking up its communicative intention. For example, some recent ERP and fMRI findings have distinguished the ERP/BOLD responses of speech gestures from other kinds of (e.g., grooming) gestures (Dimitrova, Chu, Wang, Özyürek, & Hagoort, 2016; Skipper, Goldin-Meadow, Nusbaum, & Small, 2007), the effects of gestures on the P600 syntactic ERP component are obtained only with hand gestures but not with other synchronized visual stimuli (Holle et al., 2012), and finally, some brain areas such as the left superior temporal sulcus are particularly sensitive to gesture–speech synchrony (Biau et al., 2016; Hubbard et al., 2009). Hence, even as the processing of visual (biological) motion of the gestures might partially explain the present results, we believe this interpretation would be hard to fit with data reported in several other studies on gestures.

A second interpretation of the results is that the ERP differences between gesture and no-gesture conditions reflect, at least partially, audiovisual integration. For example, Baart et al. (2014) found a modulation at N100 whenever lip movements accompanied congruent real speech or congruent sine-wave speech (before being interpreted as speech by the listeners). In contrast, the P200 component was only modulated when lip movements accompanied real speech compared to sine-wave speech before listeners interpreted it as speech. These results, together with other previous reports, suggested a multistage audiovisual speech integration process whereby early effects concurrent with N100 are associated with non speech-specific integration and a 200-millisecond time window in which P200 effects reflect binding of phonetic information. Recent studies focusing on the N100/P200 modulations to acoustic words in sentences under different conditions (e.g., speaker's accents or accompanying visual information) corroborate this interpretation of acoustic effects to N100 and phonological effects to P200 (Brunellière & Soto-Faraco, 2015; Brunellière & Soto-Faraco, 2013; Kong et al., 2010; Obleser et al., 2006). In this study, the positive shift reported in the time window corresponding to P200 may correspond to an increase of the P200 component and reflect phonetic processing triggered by the gesture's apex and the corresponding syllable (Krahmer & Swerts, 2007). In contrast, the N100 positive shift may reflect a decrease of

the N100 component and then non speech-specific effects between visual and auditory inputs. This last point might be speculative as no behavioral effect supports it in this study, and further investigation is definitely needed to address it. Additionally, as words were embodied in continuous audiovisual speech here, evoked signals were potentially noisier than in classical audiovisual phonemic studies. Thus, it is difficult to decide whether our modulations reflect an effect of gesture on audiovisual integration and/or attention by directly comparing amplitude modulations on precise components as described in the cited literature.

We think both interpretations outlined above are in fact not necessarily incompatible. In fact, even assuming that ERP modulations from gestures were limited to a simple effect of the extra visual information from the hand gesture, it is worth noting that in the case of speech gestures these effects occur systematically aligned with crucial moments in the auditory speech signal processing (i.e., aligned with critical words), as has been demonstrated many times. In any case, it would be relevant to settle the issue empirically by comparing N100/P200 modulations of word ERPs from beat gestures and visual cues without communicative intent in future investigations. Related to this issue, in an fMRI study, Biau et al. (2016) showed that BOLD responses in the left middle temporal gyrus were sensitive to the temporal misalignment between beats and auditory speech, over and above when the same speech fragment was synchronized with a circle following the original (but not present) hand trajectories.

Going back to the behavioral results, the fact that we did not find any effect of beats on the sentence interpretation can be explained by different reasons. Using the cross-split method to create the final versions of the auditory stimuli, we ensured homogenous intonations and neutrality between the nouns NP1 and NP2, as we did not want one word to be more salient compared to the other in our critical sentences. However, removing the natural prosodic breaks (pauses) after the nouns may have affected the naturalness of speech, disrupting the listeners' interpretation (as they expected to rely on auditory prosodic cues to make a decision). Nevertheless, this breach-of-naturalness account may not fully explain the null effects because the experimental sentences were screened for naturalness by the authors and three Spanish speakers with phonetic training (naïve to the experimental goals). Perhaps more likely, the lack of beat effect on sentence interpretation may be explained by the absence of the natural synergy between visual and auditory prosody. In particular, gestures might have lost their effectiveness because, as a result of the cross-splicing method to build prosody-neutral sentences, the pitch peaks normally produced by the speaker

in the stressed syllable of a word preceding a prosodic break (and hence, correlated with gestures in real speech) were removed. Previous work with pointing gestures showed that apexes and f_0 peaks are temporally aligned, and both correlated with the accented syllable (Esteve-Gibert & Prieto, 2014; this is also true for head nods: Krahmers & Swerts, 2007; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; and eyebrow movements: Krahmers & Swerts, 2007). In the present study, we might have affected the phrase final position intonations, relevant to anchor beats to acoustic envelope modulations. Recently, Dimitrova et al. (2016) showed that when a beat accompanied a nonaccented target word, it resulted in a greater late anterior positivity relative to a beat accompanying the equivalent focused word. This result actually suggested an increased cost of multimodal speech processing when nonverbal information targeted verbal information that was not auditorily emphasized. If beats' apexes and pitch peaks normally go hand in hand (Holle et al., 2012; Leonard & Cummins, 2012; McNeill, 1992), one might assume that beats lose their prosodic function when their tight temporal synchronization with acoustic anchors is affected. Thus, the absence of behavioral effect in the present study might suggest that the beats' kinetic cues alone are not sufficient to confer a prosodic value to visual information. This happened despite the fact that beat gestures produced a clear effect in terms of the neural correlates, at least the ones signaling early modulation of the auditory evoked potential. It is possible that these modulations are but one of the signals that the parsing system might use downstream to decide on the correct interpretation of a sentence. For instance, it may be relevant to look at the closure positive shift, an ERP component marking the processing of a prosodic phrase boundary and characterized by a slow positive shift observed at the end of intonational phrases (Steinhauer, Alter, & Friederici, 1999). Clear breaches in the orchestration of such signals might simply void their influence. However, this interpretation is clearly speculative and will need to be confirmed with further investigation. For example, in future investigations, it would be interesting to adapt the same procedure but maintain the temporal coherence between gesture trajectories and envelope modulations during audiovisual speech perception. It may be more sensitive to detect an incongruency effect between a beat and auditory prosodic cues on the syntactic parsing of ambiguous sentences (e.g., sentences in which a beat is aligned with NP1 but the prosodic accent is on NP2). In this manner, one could evaluate the potential emphasizing effect of beats on auditory prosody in congruent compared to incongruent conditions, in terms of hand gesture and speech envelope modulations. Alternatively, we might not have chosen a sensitive behavioral correlate to the effects of beats. Prosody does a lot of things,

and supporting syntactic parsing is only one of them. For example, because pitch also serves as a focus cue, for example, to make a word more memorable (Cutler & Norris, 1988), in future experiments it might be interesting to test postexperiment memory performance for cued words or other possibilities such as the role of gestures in ensuring attention to the audiovisual speech (e.g., lip movements and sounds). It is increasingly acknowledged that top-down attention modulates the expression of audiovisual integration in speech (Alsius, Möttönen, Sams, Soto-Faraco, & Tiippana, 2014; Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Alsius, Navarra, & Soto-Faraco, 2007;) and in general (Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010).

Finally, our exploratory analysis regarding the RC-evoked ERPs provides some initial support to the idea that gestures, indeed, produce some effect on parsing, albeit undetectable in behavior, within our paradigm. Indeed, when we looked at the P600 time window for the RC we found an amplitude reduction in the positive shift corresponding to the P600 component in both beat gesture conditions (independently from their placement on NP1 or NP2). This suggests a potential facilitation of syntactic parsing of the RC, as compared to the same sentence perceived with no gesture. Although speculative, these results are in line with what might be expected from previous literature that attached to demonstrate the syntactic processing effects neural levels. One possible speculative interpretation may be that beat gestures affect locally the saliency of the affiliate words via their role as attention cues. This local modulation may cascade onto modulations at later stages in sentence processing. These later sentence-level modulations might have been too weak in our paradigm to influence the listeners' decisions but were picked up via ERPs as a reduction in the P600 response. For instance, a study by Guellaï, Langus, and Nespors (2014) reported that the alignment of beat gestures modulates the interpretation of ambiguous sentences, which might be the behavioral counterpart of the neural signature reported in Holle et al.'s (2012) study. Further investigations are needed to fill the gap between behavioral and neural correlates.

Final revised version accepted 25 July 2017

Notes

- 1 The three measures were not available on all NPs. The analyses were performed based on the values that were available to us.
- 2 For NP1 and NP2 nouns, we also performed the same ERP analysis, adding the factor response category (HA and LA) to isolate a potential correlation between the interpretation and the effect at the N100 and P200 component time period. However, due to the limited number of epochs when separated according to the

participant's response, we had to remove more participants from the averages (in some cases, e.g., some participants adopted strategies and responded almost only HA during the whole procedure). Results did not present any significant pattern. As they were too noisy we decided that they were not reliable enough to draw any conclusion, and they have not been included.

References

- Alsius, A., Möttönen, R., Sams, M. E., Soto-Faraco, S., & Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: Evidence from ERPs. *Frontiers in Psychology*, 5, 727. <https://doi.org/10.3389/fpsyg.2014.00727>
- Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, 15, 839–843. <https://doi.org/10.1016/j.cub.2005.03.046>
- Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research*, 183, 399–404. <https://doi.org/10.1007/s00221-007-1110-1>
- Astheimer, L. B., & Sanders, L. D. (2009). Listeners modulate temporally selective attention during natural speech processing. *Biological Psychology*, 80, 23–34. <https://doi.org/10.1016/j.biopsycho.2008.01.015>
- Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, 53, 115–121. <https://doi.org/10.1016/j.neuropsychologia.2013.11.011>
- Biau, E., Morís Fernández, L., Holle, H., Avila, C., & Soto-Faraco, S. (2016). Hand gestures as visual prosody: BOLD responses to audio–visual alignment are modulated by the communicative nature of the stimuli. *NeuroImage*, 132, 129–137. <https://doi.org/10.1016/j.neuroimage.2016.02.018>
- Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, 124, 143–152. <https://doi.org/10.1016/j.bandl.2012.10.008>
- Biau, E., & Soto-Faraco, S. (2015). Synchronization by the hand: The sight of gestures modulates low-frequency activity in brain responses to continuous speech. *Frontiers in Human Neuroscience*, 9, 527–533. <https://doi.org/10.3389/fnhum.2015.00527>
- Biau, E., Torralba, M., Fuentemilla, L., de Diego Balaguer, R., & Soto-Faraco, S. (2015). Speaker's hand gestures modulate speech perception through phase resetting of ongoing neural oscillations. *Cortex*, 68, 76–85. <https://doi.org/10.1016/j.cortex.2014.11.018>
- Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer (Version 5.4.17) [Computer software].
- Brunellière, A., Sánchez-García, C., Ikumi, N., & Soto-Faraco, S. (2013). Visual information constrains early and late stages of spoken-word recognition in sentence context. *International Journal of Psychophysiology*, 89, 136–147. <https://doi.org/10.1016/j.ijpsycho.2013.06.016>

- Brunellière, A., & Soto-Faraco, S. (2013). The speakers' accent shapes the listeners' phonological predictions during speech perception. *Brain and language*, 125, 82–93. <https://doi.org/10.1016/j.bandl.2013.01.007>
- Brunellière, A., & Soto-Faraco, S. (2015). The interplay between semantic and phonological constraints during spoken-word comprehension. *Psychophysiology*, 52, 46–58. <https://doi.org/10.1111/psyp.12285>
- Carreiras, M., Salillas, E., & Barber, H. (2004). Event-related potentials elicited during parsing of ambiguous relative clauses in Spanish. *Cognitive Brain Research*, 20, 98–105. <https://doi.org/10.1016/j.cogbrainres.2004.01.009>
- Clifton, C., Carlson, K., & Frazier, L. (2002). Informative prosodic boundaries. *Language and Speech*, 45, 87–114. <https://doi.org/10.1177/00238309020450020101>
- Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human, Performance and Perception*, 14, 113–121.
- de la Cruz-Pavía, I. (2010). The influence of prosody in the processing of ambiguous RCs: A study with Spanish monolinguals and Basque-Spanish bilinguals from the Basque Country. *Interlingüística*, 20, 1–12.
- Dimitrova, D., Chu, M., Wang, L., Özyürek, A., & Hagoort, P. (2016). Beat that word: How listeners integrate beat gesture and focus in multimodal speech discourse. *Journal of Cognitive Neuroscience*, 28, 1255–1269. https://doi.org/10.1162/jocn_a_00963
- Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavioral Research Methods*, 45, 1246–1258. <https://doi.org/10.3758/s13428-013-0326-1>
- Esteve-Gibert, N., & Prieto, P. (2014). Infants temporally coordinate gesture-speech combinations before they produce their first words. *Speech Communication*, 57, 301–316. <https://doi.org/10.1016/j.specom.2013.06.006>
- Fernández, E. (2003). *Bilingual sentence processing: Relative clause attachment in English and Spanish*. Amsterdam: John Benjamins.
- Frazier, L., Carlson, K., & Clifton, C. (2006). Prosodic phrasing is central to language comprehension. *Trends in Cognitive Science*, 10, 244–249. <https://doi.org/10.1016/j.tics.2006.04.002>
- Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, 6, 78–84. [https://doi.org/10.1016/S1364-6613\(00\)01839-8](https://doi.org/10.1016/S1364-6613(00)01839-8)
- Fromont, L., Soto-Faraco, S., & Biau, E. (2017). Searching high and low: Prosodic breaks disambiguate relative clauses. *Frontiers in Language Sciences*, 8, 96. <https://doi.org/10.3389/fpsyg.2017.00096>
- Gordon, P. C., & Lowder, M. W. (2012). Complex sentence processing: A review of theoretical perspectives on the comprehension of relative clauses. *Language and Linguistics Compass*, 6, 403–415. <https://doi.org/10.1002/lnc3.347>
- Gratton, G., & Coles, M. G. H. (1989). Generalization and evaluation of eye-movement correction procedures. *Journal of Psychophysiology*, 3, 14–16.

- Grillo, N., & Costa, J. (2014). A novel argument for the Universality of Parsing principles. *Cognition*, 133, 156–187.
<https://doi.org/10.1016/j.cognition.2014.05.019>
- Guellaï, B., Langus, A., & Nespø, M. (2014). Prosody in the hands of the speaker. *Frontiers in Psychology*, 5, 700. <https://doi.org/10.3389/fpsyg.2014.00700>
- Haupt, F. S., Schlesewsky, M., Roehm, D., Friederici, A. D., & Bornkessel-Schlesewsky, I. (2008). The status of subject-object reanalyses in the language comprehension architecture. *Journal of Memory and Language*, 59, 54–96. <https://doi.org/10.1016/j.jml.2008.02.003>
- Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science*, 182(4108), 177–180.
<https://doi.org/10.1126/science.182.4108.177>
- Hirai, M., Fukushima, H., & Hiraki, K. (2003). An event-related potentials study of biological motion perception in humans. *Neuroscience Letters*, 344, 41–44.
[https://doi.org/10.1016/S0304-3940\(03\)00413-0](https://doi.org/10.1016/S0304-3940(03)00413-0)
- Holle, H., Obermeier, C., Schmidt-Kassow, M., Friederici, A. D., Ward, J., & Gunter, T. C. (2012). Gesture facilitates the syntactic analysis of speech. *Frontiers in Psychology*, 3, 74. <https://doi.org/10.3389/fpsyg.2012.00074>
- Hubbard, A. L., Wilson, S. M., Callan, D. E., & Dapretto, M. (2009). Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping*, 30, 1028–1037. <https://doi.org/10.1002/hbm.20565>
- Jahshan, C., Wynn, J. K., Mathis, K. I., & Green, M. F. (2015). The neurophysiology of biological motion perception in schizophrenia. *Brain and Behavior*, 5, 75–84.
<https://doi.org/10.1002/brb3.303>
- Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, 89, 253–260.
[https://doi.org/10.1016/S0093-934X\(03\)00335-3](https://doi.org/10.1016/S0093-934X(03)00335-3)
- Kong, L., Zhang, J. X., Kang, C., Du, Y., Zhang, B., & Wang, S. (2010). P200 and phonological processing in Chinese word recognition. *Neuroscience Letters*, 473, 37–41. <https://doi.org/10.1016/j.neulet.2010.02.014>
- Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57, 396–414. <https://doi.org/10.1016/j.jml.2007.06.005>
- Krakowski, A. I., Ross, L. A., Snyder, A. C., Sehatpour, P., Kelly, S. P., & Foxe, J. J. (2011). The neurophysiology of human biological motion processing: A high-density electrical mapping study. *NeuroImage*, 56, 373–383.
<https://doi.org/10.1016/j.neuroimage.2011.01.058>
- Lehiste, I. (1973). Phonetic disambiguation of syntactic ambiguity. *Glossa*, 7, 107–122. <https://doi.org/10.1121/1.1982702>
- Leonard, T., & Cummins, F. (2012). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26, 10.
<https://doi.org/10.1080/01690965.2010.500218>

- Marstaller, L., & Burianová, H. (2014). The multisensory perception of co-speech gestures—A review and meta-analysis of neuroimaging studies. *Journal of Neurolinguistics*, 30, 69–77. <https://doi.org/10.1016/j.jneuroling.2014.04.003>
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.
- Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, 15, 133–137. <https://doi.org/10.1111/j.0963-7214.2004.01502010.x>
- Näätänen, R. (1982). Processing negativity: An evoked-potential reflection of selective attention. *Psychological Bulletin*, 92, 605–640. <https://doi.org/10.1037/0033-2909.92.3.605>
- Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology*, 38, 1–21. <https://doi.org/10.1111/1469-8986.3810001>
- Obleser, J., Scott, S. K., & Eulitz, C. (2006). Now you hear it, now you don't: Transient traces of consonants and their nonspeech analogues in the human brain. *Cerebral Cortex*, 16, 1069–1076. <https://doi.org/10.1093/cercor/bhj047>
- Osterhout, L., & Holcomb, P. J. (1992). Event-related potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31, 785–806. [https://doi.org/10.1016/0749-596X\(92\)90039-Z](https://doi.org/10.1016/0749-596X(92)90039-Z)
- Picton, T. W., & Hillyard, S. A. (1974). Human auditory evoked potentials. II. Effects of attention. *Electroencephalography and Clinical Neurophysiology*, 36, 191–199. [https://doi.org/10.1016/0013-4694\(74\)90156-4](https://doi.org/10.1016/0013-4694(74)90156-4)
- Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of Speech, Language, and Hearing Research*, 52, 1073–1081. [https://doi.org/10.1044/1092-4388\(2009/07-0276\)](https://doi.org/10.1044/1092-4388(2009/07-0276))
- Quené, H., & Port, R. (2005). Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica*, 62, 1–13. <https://doi.org/10.1159/000087222>
- Sánchez-García, C., Alsius, A., Enns, J. T., & Soto-Faraco, S. (2011). Cross-modal prediction in speech perception. *PLoS One*, 6(10), e25198. <https://doi.org/10.1371/journal.pone.0025198>
- Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., & Small, S. L. (2007). Speech-associated gestures, Broca's area, and the human mirror system. *Brain and Language*, 101, 260–277. <https://doi.org/10.1016/j.bandl.2007.02.008>
- Steinhauer, K., Alter, K., & Friederici, A. D. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience*, 2, 191–196. <https://doi.org/10.1038/5757>
- Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, 19, 1964–1973. <https://doi.org/10.1162/jocn.2007.19.12.1964>

- Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences*, 14, 400–410. <https://doi.org/10.1016/j.tics.2010.06.008>
- Treffner, P., Peter, M., & Kleidon, M. (2008). Gestures and phases: The dynamics of speech-hand communication. *Ecological Psychology*, 20, 32–64. <https://doi.org/10.1080/10407410701766643>
- van de Meerendonk, N., Kolk, H. H., Vissers, C. T., & Chwilla, D. J. (2010). Monitoring in language perception: Mild and strong conflicts elicit different ERP patterns. *Journal of Cognitive Neuroscience*, 22, 67–82. <https://doi.org/10.1162/jocn.2008.21170>
- van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 1181–1186. <https://doi.org/10.1073/pnas.0408949102>
- Wang, L., & Chu, M. (2013). The role of beat gesture and pitch accent in semantic processing: An ERP study. *Neuropsychologia*, 51, 2847–2855. <https://doi.org/10.1016/j.neuropsychologia.2013.09.027>

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Additional Measurements.